The American Economic Review

ARTICLES

MARCH 1987

THE AMERICAN ECONOMIC ASSOCIATION

- Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

- *THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1987, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

$38.50	if annual income is $30,000 or less;
$46.20	if annual income is above $30,000, but no more than $40,000;
$53.90	if annual income is above $40,000;

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), $110 a year; individuals, $72.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $12.00 to the annual rates above to cover extra postage.

- Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

- Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

- Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

Officers

President
GARY S. BECKER
University of Chicago

President-Elect
ROBERT EISNER
Northwestern University

Vice Presidents
ANN F. FRIEDLAENDER
Massachusetts Institute of Technology
ROBERT E. LUCAS, JR.
University of Chicago

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
ORLEY ASHENFELTER
Princeton University

Managing Editor of The Journal of Economic Literature
JOHN PENCAVEL
Stanford University

Executive Committee

Elected Members of the Executive Committee
ALAN S. BLINDER
Princeton University
DANIEL L. McFADDEN
Massachusetts Institute of Technology
SHERWIN ROSEN
University of Chicago
THOMAS J. SARGENT
University of Minnesota
ROBERT J. BARRO
University of Rochester
JUDITH THORNTON
University of Washington

EX OFFICIO Members
CHARLES P. KINDLEBERGER
Massachusetts Institute of Technology
ALICE M. RIVLIN
The Brookings Institution

THE AMERICAN ECONOMIC REVIEW

•Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: $50 for members; $100 for nonmembers. Style guides will be provided upon request.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

March 1987

VOLUME 77, NUMBER 1

Articles

Shorter Papers

Erratum:

Alice M. Rivlin

Economics and the Political Process[†]

By ALICE M. RIVLIN*

I want to use this once-in-a-lifetime opportunity for pontificating to the profession, to explore ways of improving the interaction between what economists do and the political process. Tension and conflict are, of course, inherent in political decisions, especially on economic policy. Nothing can make such decisions easy. Nevertheless, it is my contention that economic policymaking in Washington in the last decade has been more frustrating, muddled, and confusing than necessary. Some of the fault lies with economists and economics; some with politicians and the political process; some in the interactions. I want to offer some suggestions for modest improvements.

Most economists probably share my premise that economics ultimately ought to be more than just challenging intellectual gymnastics. It ought to help us understand how the economy works and provide a basis for intelligent political choices among economic policies. Even those who devote their energies to resolving purely theoretical issues imagine that somehow in the end their efforts will prove socially useful.

The dedicated, idealistic young economist who aspires to advise a government may well envision herself someday as the wise and impartial adviser to the philosopher queen. In this daydream, the adviser presents the best forecasts that can be made of the future course of the economy. She explains the macroeconomic policy options and what is likely to happen if each is undertaken. She elucidates why market solutions are efficient, when markets are likely to fail, and what can be done when this occurs. She identifies risks and uncertainties, which fortunately are not overwhelming. She represents the best professional judgment of her fellow economists, indicating the major respects in which most economists agree and scrupulously pointing out that in minor respects the views of some of her professional colleagues might differ from her own. She remains above the political fray, identifying any values or distributional biases that may creep into her judgments and eschewing identification with interest groups or ideological causes.

The queen for her part listens carefully and intelligently, asks thoughtful questions, and weighs the options. She may consult other experts on noneconomic aspects of the decisions, but these can be assumed not to be very important. She then makes final decisions—even very hard ones—and sticks to them. The decisions are carried out, the economy prospers, and a grateful nation applauds the wisdom of the monarch and her economist and the usefulness of economics.

But in the real world, both economics and politics are frustratingly unlike this picture. Both are pluralistic in the extreme and appear to be getting more so. Economists and political leaders not only miscommunicate, but each accuses the other of incompetence, obfuscation, self-serving motives, and antisocial behavior.

Economists, of course, do not wait for others to attack them; they do it themselves. Walter Heller said in his presidential address that the "chorus of self criticism has risen to a new crescendo" (1975, p. 1), and the self-deprecation has not abated in the intervening decade. If a golden age of economists' self-confidence ever occurred, it is long past. Events of recent years have kept reminding us that our national economy is diverse and complex, battered by unpredictable shocks, and increasingly interconnected with the even

*The Brookings Institution, 1775 Massachusetts Avenue, N.W., Washington, D.C. 20036. The views set forth here are solely my own and do not necessarily represent the opinions of the trustees, officers, or other staff members of the Brookings Institution. I am grateful for the insights and assistance of many colleagues, especially Robert D. Reischauer, Charles L. Schultze, Mary S. Skinner, and Valerie M. Owens.

[†] Presidential address delivered at the ninety-ninth meeting of the American Economic Association, December 29, 1986, New Orleans, Louisiana.

more diverse and complex world outside our borders. Knowledge of how the domestic economy works and interacts with the rest of the world is imperfect. Economists keep coming up with ingeneous theories, but they have a hard time testing them. Data are inadequate and controlled experimentation nearly impossible. Modeling has greatly enhanced our understanding of the past, but shows few visible signs of improving the reliability of macroeconomic prediction. Forecasting even for short periods remains an uncertain art in which neither economists nor politicians can have much confidence.

Many of the most sophisticated and realistic members of the profession, conscious of all these difficulties, have abandoned the attempt to advise governments on policies in favor of the more manageable tasks of adding to the knowledge base. This may be understandable, but it deprives the economic policy debate of the input of some very good minds and runs the risk of leaving the job of interacting with the political arena disproportionately to those with strong ideological views.

I. Fragmentation of the Economic Policy Process

The pluralism of economics pales beside the pluralism of the political system that policy-minded economists aspire to assist. Even if one leaves aside the complexities of federalism, the process by which national economic policy evolves in Washington is so fragmented and complicated that it is almost impossible to explain to the uninitiated how it is supposed to work, let along how it does work.

A well-founded distrust of despots led our forefathers not only to opt for representative democracy, but to divide power among the executive and legislative and judicial branches, and between the House and the Senate. On matters of taxing and spending, they were especially protective of the power of the people's representatives, making it clear that while the president could propose taxing and spending, the ultimate authority lies with the Congress, subject only to presidential veto. This divided power creates a built-in hurdle to making and carrying out fiscal policy. The hurdle is low when the president is articulating a policy that has broad support in the country and in the Congress. It can lead to erratic shifts of policy when the president is indecisive, and to deadlock when the president is leading in a direction in which the public and its elected representatives do not wish to go. Deadlocks are rare, but can be serious. The failure to reduce the huge structural budget deficit of the mid-1980's largely reflects the fact that the president's solution—drastic reduction of the federal role in the domestic economy—does not command broad popular support.

The separation of powers between the Congress and the president is basic to our system of government and probably worth the price of occasional deadlock. The difficulties of making economic policy, however, are strongly compounded by the propensity of our pluralistic society to diffuse power and decision-making authority both within the executive branch and within Congress. With respect to taxing and spending policy, for example, the simple notion that the president proposes and the Congress disposes is greatly complicated by the fragmentation of power within each branch. Moreover, periodic efforts to make the policy process more coherent within each branch, while often temporarily successful, have added new power centers without consolidating the old ones.

In the executive branch, the trend since early in the century has been to centralize power in the White House in order to make it easier for the president to formulate and articulate taxing and spending policy, and to utilize the growing skills of the economics profession to that end. But this worthy goal has been accomplished in stages, with a new institution added at each stage. The creation of what is now called the Office of Management of Budget (OMB) in the 1920's made it possible for the president to review and evaluate spending requests and impose a set of priorities on his budget proposal to Congress reflecting his administration's view of the appropriate size and role of government. The creation of the Council of Economic Advisers (CEA) in the 1940's provided a focal point for bringing the advice of the economics profession into the service of presidential decision making and a locus for creating an official forecast of economic activity.

The creation of OMB and CEA improved the president's ability to formulate and articulate macroeconomic policy. It also left the president, in addition to his other impossible duties, with the job of resolving a built-in tension over responsibility for economic policy among the CEA, OMB, and the Treasury, not to mention the White House staff and the agencies with line responsibility for implementing various aspects of economic policy.

Presidents have tried various coordination mechanisms including "troika" arrangements and an almost infinite variety of broader councils and committees with varying membership, responsibilities, and leadership. The system works tolerably well or exceedingly creakily, depending on the president's personal style and the personalities involved. But it encourages battling over turf as well as substance, and is hardly designed to minimize the amount of presidential energy needed to evolve a coherent, explainable policy on taxing and spending. One might wonder whether it is not time to do what so many other countries do and give our president the equivalent of a responsible finance minister charged with the functions now diffused to our budget director, Council of Economic Advisers, and Treasury Secretary.

The fragmentation of power and responsibility is, of course, even more extreme in the Congress. The legislative branch also has a long history of attempts to make taxing and spending policy in a more coherent fashion by adding new coordinating institutions—appropriations committees, a joint economic committee, budget committees, a congressional budget office—without eliminating or consolidating any of the old ones.

The most recent attempt to improve congressional economic decision making—one in which I was an active participant—followed the Budget Reform Act of 1974 which created the budget committees and the Congressional Budget Office. These budget reforms succeeded in their main objective of focusing the attention of the Congress on overall budget policy, not just individual taxing and spending fragments. They have forced the Congress to fit the pieces together, to debate and vote on an overall taxing and spending plan—a budget resolution—to which specific taxing and spending matters must conform. No one can say that the Congress in the last few years has ignored fiscal policy! The creation of the Congressional Budget Office, moreover, has given Congress independent access to forecasts, projections, and analysis of economic options.

The downside of the budget reforms, however, was that the budget process was superimposed on the already complex responsibilities of authorizing, appropriating, and tax committees. It has added to the layers and stages of congressional policymaking without removing any of them, has made the process of budget decision making nearly impossible even for members of Congress to understand, and increased the workload so much that decisions are routinely made late and in an atmosphere of crisis. Moreover, Congress now frequently has to deal with two sets of estimates, those of the OMB and those of the Congressional Budget Office, which may differ because they are based on different forecasts of economic activity, or for even less obvious technical reasons.

Meanwhile, back in the separate world of the Federal Reserve, monetary policy is being decided and carried out. It is a curious paradox that a nation, which feels it needs many more hands on the tiller of fiscal policy than most countries regard as workable, is content to leave monetary policy to a central bank with fewer visible ties to the rest of the government than the central banks of most countries.

There is plenty of informal communication, of course, especially between the Federal Reserve and the hydraheaded economic establishment of the executive branch. More formal cooperation between the monetary and fiscal authorities, as in the United Kingdom, might contribute only marginally to making monetary and fiscal policy decisions part of a more coherent strategy for the economy—and at the cost of depriving the executive branch of the luxury of blaming the Federal Reserve when things go wrong. The love-hate relation between the Congress and Federal Reserve, however, warrants more attention. Despite occasional outbursts of anxiety over escalating interest rates, Congress has shown little inclination to control monetary policy, or even to inquire into the consistency of monetary and fiscal objectives. The Fed is required to report monetary growth targets to the banking committees, as though monetary policy were a matter of banking system regulation, but has little genuine interaction with the budget committees whose job is to debate and propose fiscal policy.

II. The Process under Stress

This whole complicated economic policy system has been subjected to enormous strain in recent years. Political economists like to harken back to the golden years of the 1950's and 1960's when economists got respect and the economic policy machinery functioned smoothly. The nostalgia is only partly a result of faulty memories. It's not hard to be satisfied with economists and policy processes when the economy is growing, productivity marches steadily upward, and even the national debt is obligingly declining in relative importance. It's much harder when productivity growth plummets for reasons that no one honestly purports fully to understand, expectations of public and private consumers have to be cut back to fit with slower income growth, and inflation and interest rates are bouncing around at unfamiliar levels.

Adjusting to the energy shocks and slower growth that began in the 1970's strained the economic policy processes of all industrial countries and made the participants feel frustrated and inadequate. It's not obvious, even with hindsight, that the fundamental difficulties facing the industrial world in the 1970's can credibly be blamed on economists or any particular structure of government or economic policy responses, but all came in for their share of the understandable hostility.

The difficulties of the U.S. economy in the 1980's, by contrast, revolve heavily around an economic policy mistake: the creation of a large structural deficit in the federal budget. I do not believe that the structure of our economic decision process was the cause of the mistake. Blaming the deficit on inherent flaws in the policy process requires an explanation of why the process did not cause similar mistakes in the past. But the events of 1981 which produced the deficit illustrate several of the difficulties of economic policymaking which make mistakes harder to avoid:

the uncertainty of macroeconomic forecasting;

the isolation of monetary and fiscal policy;

the contentiousness of economists and their tendency to let their ideological positions cloud their judgments about the likely effects of particular policies.

That a tax cut unmatched by comparable spending cuts would produce a deficit should have surprised no economist. That the deficit was so large reflected both economic and political miscalculations. The Reagan Administration has been faulted for masking the deficit with a "rosy scenario," but the fact is that most of the forecasting community, including the Congressional Budget Office, expected positive real growth in the economy. The administration's official forecast differed from the rest only in its degree of optimism. Forecasters in and out of government were oversanguine about growth largely because they failed to realize how serious the Federal Reserve was about reining in the money supply to control inflation. The Fed was not defying the administration, which was touting the efficacy of monetary stringency for controlling inflation, but hardly anyone seemed to remember that the way tight money controls inflation is by slowing economic activity. Moreover, as our Association's President-elect, Robert Eisner, has pointed out (1986, p. 146), the economics community, unfamiliar with a world of high inflation rates, overestimated the stimulative effect of the existing deficit. Added to this was the enthusiasm of the ideological proponents of smaller government, some of whom exaggerated the possible effects of lower tax rates on supply and some of whom simply hoped that deficits would pressure Congress to cut back domestic spending. The size of the deficits was also masked by the assumption of unspecified future spending cuts, an assumption reflecting the view that the U.S. government was operating a lot of wasteful programs with little public support which Congress could soon be persuaded to reduce or eliminate.

Both in the administration and in Congress, decisions were made at a breakneck pace, in a highly charged political atmosphere, amid conflicting claims and competing forecasts, with little attention to the consistency of monetary and fiscal policy and mostly by people with little experience in evaluating the reasonableness of any set of economic estimates. (See David Stockman, 1986, ch. 3.) When the dust settled, we found ourselves with a serious recession that nobody expected, and an escalating structural budget deficit that nobody wanted. It was hardly economic policy's finest hour.

The agonizing—and so far only partially successful—struggle to correct the mistakes of 1981 have kept the economic policy process under stress and have continued to dramatize some of its weakest aspects. The struggle between the president and the Congress over deficit solutions illustrates the price we pay for the separation of powers. The fact that fiscal policy has become an exercise in damage control, while the Federal Reserve makes all the important decisions about the economy, underlines the separation of monetary and fiscal policy. The sensitivity of deficits to the pace of the economy advertises the unreliability of macroeconomic forecasts. The fact that all the actions that could be taken to correct the deficit are unpleasant ones drags out the annual agony of budget setting interminably and dramatizes how layered and cumbersome it has become.

Small wonder that the strains of the last few years, with a little help from the press, have reinforced the negative stereotypes that economists and political decision makers have of each other. Political decision makers see economists as quarrelsome folks who cannot forecast, cannot agree, cannot express themselves clearly, and have strong ideological biases. Economists return the favor by regarding politicians as shortsighted, interested only in what is popular with the electorate, and unwilling to face hard decisions. All of the stereotypes are partly right.

Politicians embody their stereotype in economist jokes. Economists have retaliated more massively by applying the tools of their trade to the political system itself. Public choice theory essentially asks the question: what would economic policy be like if our stereotype of politicians were entirely true? The answer provides considerable insight into observed political behavior and certainly helps explain why the idealistic economist so often fails to find the system simulating the public interest motivation of the philosopher queen.

III. Some Drastic Nonsolutions

Widespread concern that the economic policy process is not working well has spawned proposals for drastic change that move in two quite different directions: one toward circumscribing the discretion of elected officials by putting economic policy on automatic pilot and the other toward making elected officials more directly responsible to the voters for their policies.

The automatic pilot approach flows from the perspective of public choice theory that the decisions of democratically elected officials interested in staying in office cannot be counted on to produce economic policy in the social interest, but are likely to be biased toward excessive government spending, growing deficits, special interest tax and spending programs, and easier money. A way to overcome these biases is to agree in advance on strict rules of economic policy, such as a fixed monetary growth path or constitutionally required balance in the federal budget.

Even if one accepts the premises, however, firm rules are hard to define in a rapidly changing world—no one seems to know what "money" is anymore—and can easily lead to perverse results. Recent experience with trying to reduce the federal deficit along the fixed path specified by the Gramm-Rudman-Hollings amendment, for example, has given us a taste of some of the possible disadvantages of a balanced budget rule. There is danger that specific dollar targets for the deficit will require procyclical fiscal policy, perhaps precipitating a recession that would then make budget balance even less attainable. Moreover, the effort to reach the targets can induce cosmetic or self-defeating measures, such as moving spending from one fiscal year to another for no valid reason, selling assets to reduce a current deficit while exacerbating future ones, and accomplishing desired purposes by regulatory or other nonbudgetary means.

The Gramm-Rudman-Hollings experience, however, has suggested the usefulness of a different approach to deficit reduction than a balanced budget rule; namely, a deficit neutral amendment rule. If legislators advocating a tax preference are required to propose a rate increase to pay for it, special interest tax legislation may falter. Similarly, the requirement that a proposal for additional spending be accompanied by a simultaneous proposal to raise taxes or reduce another spending program may be an effective brake on deficits.

The other direction of reform reflects the contrasting view that the separation of powers and the diffusion of responsibility in our government make it too difficult for the electorate to enforce its will by holding officials responsible for their policies. The potential for deadlock would be reduced if the United States moved toward a parliamentary system, or found a way to hold political parties more strictly accountable for proposing or carrying out identifiable policies.

Casual examination of parliamentary democracies, such as the Untied Kingdom and Sweden, does not provide striking evidence of the superiority of parliamentary systems for making economic choices, even if one did not have two hundred years of tradition to contend with in changing our system. The more modest notion that our system would work more smoothly if political parties had better defined positions and disciplined their elected members more strictly may well be right, but seems to fly in the face of current history. Voters are showing less strong party affiliation and more inclination to choose for themselves among candidates, while members of Congress tend increasingly to be pragmatists willing to work out nonideological compromises across party lines. These trends seem likely to be the irreversible consequences of greater education, sophistication, and exposure to public issues among voters and elected officials alike and to make a resurgence of party discipline and loyalty unrealistic.

IV. Making the Economic Policy System Work Better

My own proposals involve less drastic changes in the structure of our government. They reflect a strong faith in the ability of informed citizens and their elected representatives to make policy decisions for the common good, even to make substantial sacrifices and take political risks to further what they perceive as the long-run national interest—once they understand what the choices are. I also believe that the separation of powers between the executive and legislative branches works pretty well most of the time. It provides needed protection against overzealousness in either branch, albeit at some risk of occasional stalemate.

The main problem, it seems to me, is that our economic policy system has gradually become so complex, diffused, and fragmented that it impedes rather than fosters informed choices on major issues. The fragmentation imposes two kinds of costs. First, it makes the decision process itself exceedingly inefficient. Decisions are made too often, in too great detail, and reviewed by too many layers of decision makers in the executive branch and in Congress. Too much time is absorbed in procedure and in wrangling over details, not enough on major decisions. It's time to simplify the process, to weed out some of the institutions, and to tip the balance between substance and process back toward substance.

Second, decisions are made separately that ought to be made together, or at least with attention to their impact on each other. The separation of monetary and fiscal policy is one example; the separation of tax and spending decisions is another. Congress has made a good deal of progress in recent years in putting spending decisions together with their revenue or deficit consequences, but more could be done. I have seven steps to suggest that might make the economic policy process work more effectively.

First, seek out decisions that should be made less frequently and arrange to do so. This would economize decision-making time and enhance the chances of thoughtful, well-informed decisions. It would free up time and energy for managing the government enterprise more effectively, with a longer planning horizon. It would also reduce the inefficiency and sense of unfairness that goes with frequent changes of the rules. Making the federal budget every other year would be a major advance. Major revisions of the tax code should occur even less frequently. Big ticket acquisitions, such as major weapons systems, should be reviewed thoroughly at infrequent intervals and then put on a steady efficient track, not constantly revisited.

With a two-year budget, there would occasionally be major events, such as a sudden escalation of international tension or a sharp unexpected shift in the economic outlook, that would justify reopening the budget in midstream, but the temptation to tinker frequently should be strongly resisted. The argument that economists cannot forecast accurately two years in advance, while quite true, does not undermine the case for a multiyear budget. It simply reinforces the point that discretionary fiscal policy is hazardous and ought to be viewed with great skepticism whether the budget is annual or biennial.

Second, seek out decisions that need not be made at all and stop making them. Some spending programs could be consolidated into block grants or devolved to the states, not necessarily in the interest of smaller government, but in the interest of greater responsiveness to local needs and a less cluttered federal decision schedule. In other cases, the responsibility is clearly federal—as in defense—but Congress would be doing its job more effectively if it concentrated on major policy issues rather than on details of program management.

Third, in the executive branch, consolidate authority for tax, budget, and fiscal policy in a single cabinet department. The department could retain the name Treasury, but might better be called the Department of Economic Affairs. The Secretary of Economic Affairs should have a high level chief economist or economic council with a strong professional staff. The chief economist should work closely with the budget director who also should report to the Secretary. The purpose would be to bring together economic decisions now made in OMB, CEA, and Treasury under one high-level responsible person, to relieve the president of the duty of adjudicating among so many potentially warring power centers, and to increase the chances of building a highly professional permanent economic staff one step removed from the short-run political concerns of the White House.

Fourth, streamline the congressional committee structure to reduce the number of steps in the budget process. The authorizing and appropriating functions should be combined in a single set of "program committees," one for each major area of public spending. This would imply a single defense committee, for example, and a social insurance committee. The tax committees should handle the revenue side—not additional spending programs as at present. The budget committees would be charged with considering fiscal policy and putting the spending and revenue sides together into a budget to be passed by the whole congress. The Joint Economic Committee should celebrate the important contributions it made to economic understanding in the days before the budget process and then close up shop.

Fifth, bring monetary and fiscal policy into the same conversation. This end could be furthered by closer formal links between the central bank and the Department of Economic Affairs to dramatize the need for consultation and interaction. The Federal Reserve chairman should make a report to the budget committees of Congress laying out recommended short- and longer-run economic goals for the nation and discussing combinations of monetary and fiscal strategies to achieve them. The Fed's report should be an important input to congressional deliberations on fiscal policy.

Sixth, strive for a government-wide official economic forecast to be updated on a regular schedule. The main purpose of the common forecast would be to reduce the confusion generated by conflicting estimates, but the increased interaction between the Department of Economic Affairs, the Congressional Budget Office, and the Federal Reserve necessary to create such a forecast would increase mutual understanding of what is happening to the economy and what the goals of policy should be. Occasionally, it might be necessary for one of the agencies to dissent and explain why it disagreed with the forecast, but these occasions are likely to be infrequent. There should also be more attention than at present to the consequences for policy of the forecast being wrong.

Finally, bring choices explicitly into the decision process, both in executive branch deliberations and, especially, in Congress. Those proposing spending increases or tax reductions should routinely be required to specify what is to be given up and to offer both the benefit and its cost as a package. In other words, proposals should be deficit neutral.

V. What Economists Can Do

For their part, how can economists be more useful in the policy process? The press and politicians often sound as if they are telling us to work harder: go back to your computers and don't come out until you known how the economy really works and can give us reliable forecasts. But economists know that the economic system is incredibly complicated, and that increasing global interdependence and rapidly changing technologies and public attitudes are not making it easier to understand. It is not likely in our lifetimes that anyone will happen on a paradigm that explains everything, or even that forecasting will become appreciably more accurate. Like the medical profession, which also deals with an incredibly complex system, we economists just have to keep applying our imperfect knowledge as carefully as possible and learning from the results. Both doctors and economists need humility, but neither should abandon their patients to the quacks.

The objective of economists ought to be to raise the level of debate on economic policy, to make clear what they know and do not know, and to increase the chances of policy decisions that make the economy work better. Much of the time that means telling the public and politicians what they would rather not hear: hard choices must be made. We are stuck with being the dismal science.

Increased effort in three directions would make economics more useful in the policy process. First, *economists should put much more emphasis on their areas of agreement.* The press admittedly makes this difficult. Agreement is not news, and the press' stereotype of economists' diversity of views is so entrenched that they will go to great lengths to scare up a lonely dissenter to an almost universally held economic platitude and give her equal time.

Economists realize that the breakthrough insights around which "schools" are built are at best partial visions of the truth, but our training leads us to elaborate and differentiate these insights, to explain to ourselves and to others where they lead in different directions, not where they come together. Yet areas of agreement are wide—even in macroeconomics—and a major effort to make this clearer to ourselves and our audience would be useful.

Second, *economists should devote more serious attention to increasing the basic economic literacy of the public, the media, and the political community.* While the print media seem to me increasingly knowledgeable and sophisticated about economic issues, television, where most people get most of their information, lags far behind. Television coverage of the economy is heavily weighted to isolated economic statistics reported without context—the wholesale price index increased two-tenths of a percent in October—and talking heads disagreeing, briefly, for some obscure reason. Some of the best newscasters appear to have bad cases of economics phobia.

Media bashing is not the answer. The profession needs to take the lead in explaining more clearly what is happening to the economy, why it matters, and what the arguments are about or ought to be about. This means more than each of us taking a little time to make a luncheon speech, write an op ed piece, or appear on a talk show. It means sustained efforts on the part of teams of economists to figure out how to present economic ideas more interestingly and understandably, developing new graphics and other teaching tools and getting feedback from real audiences. The technology is available and the audiences exist—the number of people who will watch long hard-to-follow congressional debates and hearings on cable television is quite astonishing. We just need to devote the kind of effort and ingenuity that goes into explaining to audiences the complex, fast-moving, jargon-ridden game of football to our complex, fast-moving, jargon-ridden game of economics.

Third, *economists need to be more careful to sort out, for ourselves and others, what we really know from our ideological biases.* George Stigler pointed out in his presidential address (1965) that economists beginning with Adam Smith have not hesitated to make strong assertions, both positive and negative, about the effectiveness of government intervention without offering serious evidence to support their claims. For two hundred years, "the chief instrument of empirical demonstration on the economic competence of the state has been the telling anecdote" (pp. 11–12). In the more than two decades since Stigler presided over our Association, an enormous amount of useful empirical work has been done, as he predicted it would be, on the effectiveness of government programs, the costs and benefits of regulation, and so forth. Still the arguments among economists about the merits of larger vs. smaller government too often revolve around anecdotes or, worse, misleading statistics quoted out of context. My own anecdotal evidence would lead me to believe that liberals and conservatives are about equally guilty.

My concern is not with economists taking sides on policy issues or acting as advocates of particular positions. Indeed, I think many policy debates would be clarified if there were more formal and informal opportunities for economists to marshall the evidence on each side and to examine and cross-examine each other in front of some counterpart of judge or jury.

We economists tend to be uncomfortable in the role of partisans or advocates, preferring to be seen as neutral experts whether we are or not. Lawyers move more easily among roles; and the best are able to serve with distinction at different times as prosecutors, defenders, experts, and judges. The system works well when the roles are played competently and the rules of evidence strictly observed. Economists might increase their usefulness to the policy process if they made clear at any given moment which role they were playing. More important, we need to work hard to raise the standards of evidence, to make clear to the public and the participants in the political process what we are reasonably sure we know and how we know it, and where we are guessing or expressing our preferences.

REFERENCES

Eisner, Robert, *How Real is the Federal Deficit*, New York: Free Press, 1986.

Heclo, Hugh, "OMB and the Presidency—the Problem of 'Neutral Competence'," *The Public Interest*, 1975, *10*, 80–89.

Heller, Walter, W., "What's Right With Economics," *American Economic Review*, March 1975, *65*, 1–26.

Mueller, Dennis C., *Public Choice*, Cambridge: Cambridge University Press, 1979.

Okun, Arthur M., "The Economist and Presidential Leadership," in *Economics for Policy Making*, Joseph A. Pechman, ed., Cambridge: MIT Press, 1983, 577–82.

Porter, Robert B., "Economic Advice to the President: From Eisenhower to Reagan," *Political Science Quarterly*, Fall 1983, *98*, 403–06.

______, "Organizing Economic Advice to the President: A Modest Proposal, "*American Economic Review Proceedings*, May 1982, *72*, 356–60.

Schultz, George, "Reflections on Political Economy," *Challenge*, March/April 1974, *17*, 6–11.

Schultze, Charles L., "The Role and Responsibilities of the Economist in Government," *American Economic Review Proceedings*, May 1982, *72*, 62–66.

Stigler, George J., "The Economist and the State," *American Economic Review*, March 1965, *55*, 1–18.

Stockman, David A., *The Triumph of Politics*, New York: Harper and Row, 1986.

Tufte, Edward R., *Political Control of the Economy*, Princeton: Princeton University Press, 1978.

On the Marginal Welfare Cost of Taxation

By EDGAR K. BROWNING*

This paper develops a rigorous partial-equilibrium analysis of the determinants of the marginal welfare cost (MWC) of taxes on labor earnings. It shows that four key parameters interact to determine the magnitude of MWC. Using aggregate data and plausible ranges of values for the parameters, MWC can vary from under 10 percent to more than 300 percent of marginal tax revenue, suggesting that, given available evidence, we cannot estimate MWC with much precision.

The marginal welfare cost of raising tax revenue is now understood to be an important factor in the analysis of government expenditure policies, and several recent studies have developed estimates suggesting its size is substantial.[1] In general, these studies have concluded that the marginal welfare cost is significantly larger than I found in my early study (1976). For example, I concluded that marginal welfare cost was likely to be between 9 and 16 percent of additional revenue raised, but Charles Ballard, John Shoven, and John Whalley (1985) suggest that it is in the 15 to 50 percent range, with Charles Stuart (1984) reporting similar results. Developing an analysis that clarifies why the estimates differ so markedly is a major purpose of the present paper.

Both Stuart and Ballard et al. employ general-equilibrium methodologies, while I used a simple partial-equilibrium formulation based on Arnold Harberger's (1964) approach. It is apparently widely believed that this difference in methodologies is responsible for the difference in results, with the general-equilibrium approaches capturing some essential elements that are missing in the partial-equilibrium approach. I do not believe that this is the case; almost all of the differences in results can be traced to different assumptions about key parameter values. To support this assertion, this paper develops the partial-equilibrium approach in a more careful and usable form, and shows that modest variations in four key parameters can account for much of the apparent differences in results. One of the virtues of the partial-equilibrium approach is that it clarifies the contribution these key parameters make to the final estimate, something that is often obscured in large-scale general-equilibrium models.[2]

Section I develops the theory necessary to estimate the total welfare cost due to labor supply distortions of the tax system. It also corrects an error in the original Harberger formulation that I used, which led to an underestimate of total and marginal welfare costs in my 1976 paper. Section II applies the theory to the calculation of the marginal welfare cost of raising tax revenue and shows that by varying four parameter values over a relatively narrow range, the estimated marginal welfare cost varies from under 10 percent to well over 100 percent.

I. The Total Welfare Cost

Here I will consider only the welfare cost that results from taxes on labor incomes, both because the theory is less controversial and because there is a greater consensus

*Department of Economics, Texas A&M University, College Station, TX 77843. I thank Charles Ballard, Donald Deere, Charles Stuart, and anonymous referees for comments on previous drafts.

[1] See Charles Ballard et al. (1985), Ingemar Hansson and Charles Stuart (1983), Stuart (1984), and David Wildasin (1984).

[2] This is especially true in the case of Ballard et al. (1985), where the model is a multisector, dynamic computational general-equilibrium model. On the other hand, the far simpler two-sector general-equilibrium model of Stuart does a better job of focusing on the importance of key parameter values. The model in the present paper is more in the spirit of Stuart's approach.

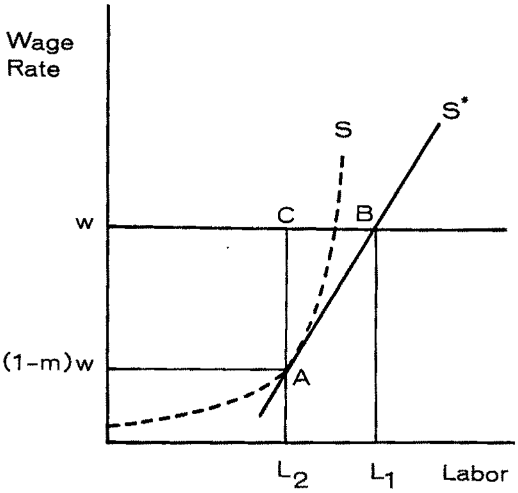


FIGURE 1

concerning empirical magnitudes than for taxes that fall on capital income. Figure 1 illustrates the usual representation of the welfare cost that results from a tax on labor income. The worker's wage rate is w (assumed to equal the marginal value product of his labor services), and labor earnings are subject to a tax at a marginal rate of m, so that the net marginal wage rate confronting the worker is $(1-m)w$. The equilibrium in the presence of the tax is at point A, where the quantity of labor supplied is L_2.[3] The compensated labor supply curve drawn for the utility level realized by the worker with the tax in place is S^*.[4] (Ignore supply curve S for the moment.) Thus, the total welfare cost is shown by area ACB, equal to the increase in earnings if the marginal tax rate is reduced to zero (but with the worker kept on the same indifference curve), CBL_1L_2, less the value of leisure given up in generating that increment in earnings, ABL_1L_2.[5]

It is important to recognize that area ACB is an exact measure of the welfare cost of the tax within the context of this model; there is no approximation involved. The key point is that I am using the compensated labor supply curve, which is necessary when evaluating welfare effects of changes in labor supply. However, note that this analysis is based on the assumption that the market wage rate remains unchanged when labor supply changes from L_2 to L_1. This assumption, common but not essential in partial-equilibrium models, differs from the general-equilibrium treatment in which the market wage rate is endogenously determined. The appropriateness of the fixed wage rate assumption will be discussed later.

To derive a formula that can be used to calculate the total welfare cost, it is assumed that the compensated labor supply curve is linear between L_2 and L_1. Then the welfare cost, W, equals one-half $CB \times AC$, or,

$$W=\tfrac{1}{2}(dL)wm. \tag{1}$$

The compensated change in the quantity of

[3] It is important to understand that the line between $(1-m)w$ and A in Figure 1 should not be interpreted to mean that the marginal tax rate is necessarily constant regardless of the level of earnings. The welfare cost depends on the marginal tax rate at the actual earnings level, which is identified here as m; the marginal tax rate(s) that applies to inframarginal earnings may differ from this. Thus, Figure 1 should not be taken to imply a proportional tax, but only to emphasize that it is the tax rate at the margin (evaluated at the worker's actual equilibrium position) that produces the distortion in the allocation of resources. In particular, note that if the tax is progressive, tax revenue will not be equal to the rectangle $wCA(1-m)w$; it will be smaller than this because the marginal tax rate that applies to earnings below wL_2 is less than m.

[4] This compensated supply curve is drawn for the utility level realized by the worker after adjustment to the tax and whatever benefits are received from government expenditures. Government expenditures are held constant along the compensated supply curve.

[5] Although Peter Diamond and Daniel McFadden (1974) have proposed a different measure of welfare cost, I believe this continues to be the standard measure. Put differently, area ACB is equal to the difference between the tax revenue actually collected and the revenue that could be collected with a lump sum tax that leaves the taxpayer on the same indifference curve that he attains under the actual tax. This is equivalent to the measure defended by J. A. Kay (1980) in his criticism of Diamond and McFadden; Kay describes the measure as the difference between tax revenue and the equivalent variation measure of the loss in consumers' surplus from the tax. By contrast, the Diamond-McFadden measure uses the compensating variation measure of the change in consumers' surplus and the tax revenue that would hypothetically be collected at the compensated equilibrium.

labor supplied can be expressed as the inverse of the slope of the compensated supply curve, dL/dw, times the change in the marginal wage rate, wm, so

$$(2) \qquad W = \tfrac{1}{2}\left[\frac{dL}{dw}wm\right]wm.$$

Multiplying by $L_2(1-m)/L_2(1-m)$ yields

$$(3) \quad W = \tfrac{1}{2}\left[\frac{dL}{dw}\frac{w(1-m)}{L_2}\right]\frac{m^2}{1-m}wL_2.$$

Note that the term in brackets equals the elasticity of the compensated supply curve evaluated at the net of tax wage rate (point A in Figure 1). Expressing this compensated labor supply elasticity as η, equation (3) can be conveniently written as[6]

$$(4) \qquad W = \tfrac{1}{2}\eta\frac{m^2}{1-m}wL_2.$$

In contrast to equation (4), the widely used Harberger formula for calculating the welfare cost is

$$(5) \qquad W = \tfrac{1}{2}\eta m^2 wL.$$

It is easily shown that the Harberger formula correctly evaluates the welfare cost if we measure the compensated elasticity and the level of labor earnings at their undistorted levels, that is, at point B in the diagram. However, these values are not observable, and available estimates pertain to elasticities and earnings evaluated in the presence of distorting taxes, that is, at point A in the diagram. Consequently, equation (4) will generally be the appropriate way to estimate the total welfare cost of a tax on labor earnings.

In my earlier paper (1976), I started with (Harberger's) equation (5) and from it developed expressions to estimate the marginal welfare cost. This procedure led to an underestimate of total and marginal welfare costs; my earlier estimates should be multiplied by (approximately) $1/(1-m)$ to correct for this error. This is one reason why recent general-equilibrium studies have generally found larger welfare costs—an error in my use of, rather than a true shortcoming of, the partial-equilibrium approach.[7] I avoid this error here by not relying on the Harberger formula.

Before turning to the issue of marginal welfare cost, it will be helpful to consider the application of this approach to the estimation of the total welfare cost, in part because this clarifies several points that are also relevant for the estimation of marginal welfare costs. For this purpose, I propose to use equation (4) with aggregate rather than individual data. If all households confronted the same marginal tax rate and had the same labor supply elasticity, this approach would yield the correct result. However, as can easily be shown, when marginal rates and/or elasticities differ, this common approach understates the welfare cost, and the understatement is larger the greater the dispersion in marginal tax rates and elasticities. Although I do not believe the actual dispersion is large enough to greatly affect the estimates (at least relative to the other factors I wish to emphasize here), the downward bias of this approach should be kept in mind.[8]

[6] Note that the average tax rate does not enter into the determination of the welfare cost according to equation (4). However, this does not mean that the average tax rate plays no role; it can influence the welfare cost through its indirect effect on the labor supply elasticity and earnings. For example, for an unchanged marginal tax rate, a higher average tax rate will increase the wL term if leisure is normal, and since the worker ends up on a different indifference curve, the compensated supply elasticity may also be affected. To apply equation (4) correctly, we do not need to know the average tax rate, but we do need to know the compensated supply elasticity and earnings at the worker's actual equilibrium position, thereby incorporating whatever effect the average tax may have through these terms.

[7] This error in the use of the Harberger formula has been pointed out in Christopher Findlay and Robert Jones (1981).

[8] Jerry Hausman (1981) uses disaggregated data in his work estimating welfare costs. Potentially, this approach will yield more accurate estimates, but there are some serious problems with his implementation of this approach (see my 1985b paper), and he does not provide estimates that permit a comparison of the dif-

To apply equation (4), we require estimates of aggregate labor earnings, a weighted-average compensated labor supply elasticity for workers as a group, and a weighted-average marginal tax rate for workers as a group. Although the greatest uncertainty surrounds the appropriate value for the labor supply elasticity, there is no point in reviewing once again the econometric literature, and I will simply use values of 0.2, 0.3, and 0.4 here. While values substantially larger than 0.4 have been used in the literature, it seems unlikely to me that a value much in excess of this figure is plausible.[9]

The only subtle point to recognize in choosing a value for aggregate labor earnings is that labor supply should be valued at the marginal value product of labor since the theory is based on the tax wedge between the marginal value product and the net wage received by workers. (See Figure 1 where *w* is the marginal value product.) In the absence of indirect taxes collected from firms (and some other factors mentioned below), wage earnings received by workers would represent the appropriate magnitude. However, because of the employer portion of the Social Security payroll tax, fringe benefits, and indirect output taxes (sales and excise taxes), reported wage and salary incomes must be grossed up to a broader measure of before-tax labor compensation. A rough estimate of the required figure for 1984 is $2400 billion.[10] This compares with wage and salary income of only $1800 billion.

The weighted-average marginal tax rate should reflect the combined effect of all taxes and transfers in reducing the net marginal wage rate received by workers below the marginal value product of labor. Thus, the marginal tax rate should be measured relative to the broad before-tax measure of labor income. This means that statutory tax rates are not the appropriate values to use. To see this, consider the Social Security payroll tax which was levied at a 14.1 percent combined employer-employee rate in 1984. If a worker increases his labor supply sufficiently to receive an additional $100 from his employer, he actually had to generate $107.05 in additional product since the employer portion of the tax ($7.05) is remitted to the government before the worker is paid. Thus, the marginal tax rate that applies to the worker's marginal value product is $14.10/$107.05, or 13.2 percent rather than 14.1 percent (assuming no other indirect taxes, fringe benefits, and so on).

Similarly, the effective marginal tax rate of personal income taxes is below the statutory

ference when aggregate data are used. An example of how sensitive the results are to dispersion in marginal tax rates is provided by the following. Consider three workers with respective earnings of $10,000, $20,000, and $30,000, who confront marginal tax rates of 30, 37, and 44 percent, respectively. With a compensated labor supply elasticity of 0.3, using equation (4) with the individual data and summing yields an estimate of $2400 for the total welfare cost. Using aggregate data—$60,000 for earnings, 0.3 for the elasticity, and the weighted-average (weights equal to share of total labor income) marginal tax rate of 39.4 percent—the estimate is $2305, only 4 percent less than the correct figure. Of course, the difference will be larger if the differences in marginal tax rates are greater. However, my paper with William Johnson (1984, Table 3) found that the average effective marginal tax rates for the top four quintiles of households range only from 39 to 47 percent when all taxes and implicit marginal tax rates of transfers are taken into account. Of course, there is also variation in marginal tax rates within quintiles, so the degree of understatement may be larger than these figures suggest.

[9] Numerous references to the relevant literature are contained in my paper with Johnson, Ballard et al. (1985), and Stuart. It should be noted that both Stuart and Ballard et al. use upper bound values for the compensated labor supply elasticity that exceed the 0.4 figure used in this paper. Stuart uses a value of 0.836; while Ballard et al. do not explicitly give the value they use, based on Table 1 of Ballard et al. (1982), the figure is apparently about 0.6. These figures seem too high to me, although there is some empirical evidence to support such values. Note that with a marginal tax rate of 43 percent, a compensated labor supply elasticity of 0.6 implies that reducing the marginal tax rate to zero in a compensated fashion would increase labor supply by 45 percent.

[10] The *Economic Report of the President* (1985, Table B-21) gives total compensation of employees (which includes the employer contribution to Social Security and some fringe benefits) for 1984 as $2173 billion. To this can be added the approximate $147 billion in sales and excise taxes which, according to M. Kevin McGee (1985), can be taken to fall on labor income. In addition, I assume that $80 billion of the $155 billion in proprietors' income represents labor compensation.

marginal tax rate that applies only to taxable income as defined by the tax laws. The significance of this point is evident from a comparison of the results of recent studies by Robert Barro and Chaipat Sahasakul (1983) and John Seater (1984). Barro and Sahasakul estimate a weighted-average marginal tax rate in 1980 for the federal individual income tax of 30.4 percent; this is simply an average of statutory marginal tax rates weighted by adjusted gross income. For the same year, Seater estimated a weighted-average marginal tax rate of 22.2 percent, but he arrived at his estimate by relating actual tax payments to variations in adjusted gross income (rather than to taxable income). For purposes of evaluating the labor supply distortions of taxes, the Seater approach comes closer to measuring the effective marginal tax rate that applies to the marginal value product of labor.[11]

In addition to measuring each tax's effective marginal tax rate consistently with respect to the same broad base, it is the combined marginal tax rate due to all factors that depress the marginal net wages received by workers that is relevant. Thus, the implicit marginal tax rates of means-tested transfer programs must also be included. One study that does measure marginal tax rates due to all taxes and transfers relative to a broad measure of income is my paper with William Johnson, which provides estimates for each quintile of households for 1976. A weighted average (weights equal to each quintile's share of labor income, broadly measured) of these marginal tax rates is 43 percent, and I will use this as my benchmark estimate for the effective marginal tax rate in 1984.[12]

There are, however, greater difficulties involved in accurately estimating the effective marginal tax rate than are commonly recognized, and the 43 percent figure should be viewed as subject to a significant margin for error.[13] For example, the Browning-Johnson estimate, as well as most others, treats the Social Security payroll tax as fully a distortion at the margin (except for those earning above the ceiling on taxable earnings). But if workers view, correctly or not, an additional dollar in Social Security taxes as purchasing deferred labor compensation in the form of a pension with a present value of a dollar, then the effective marginal tax rate of this tax would be zero.[14]

In view of this consideration, as well as others, it is appropriate to consider a range of values for the weighted-average effective marginal tax rate. Consequently, I use values of 38, 43, and 48 percent in the calculations. These estimates, together with the compensated labor supply figures (0.2, 0.3, and 0.4) and gross labor compensation ($2400 billion), can be inserted into equation (4) to estimate the total welfare cost of distorted labor supply decisions in 1984.

Table 1 displays the results, with the total welfare cost as a percentage of tax revenues from taxes that fall on labor income shown in parentheses.[15] What is perhaps most strik-

[11] To the extent that some exclusions and deductions are worth less at the margin than after-tax cash income, the approach used by Seater would understate the effective marginal tax rate to some degree.

[12] The Browning-Johnson estimate for 1976 is really a weighted-average marginal tax rate for labor and capital taxes together as they apply to an increment of labor and capital income. Insofar as the marginal tax rate on labor income is lower than the marginal tax rate on capital income, this figure would overstate the rate on labor income. However, since 1976, labor income has come to be taxed more heavily.

[13] See myself and Johnson, Barro-Sahasakul, and Seater for discussions of some of the technical problems.

[14] Three recent studies have investigated the linkage between social security taxes and future benefits (Roger Gordon, 1983; myself, 1985a; and Richard Burkhauser and John Turner, 1985), but with conflicting results. It seems quite possible, however, that the effective marginal tax rate of Social Security is somewhat less than the approximate 9 percentage point contribution it makes to the overall 43 percent rate cited above.

[15] Total tax revenues from taxes on labor income in 1984 are approximately $745 billion. This is the sum of Social Security payroll taxes ($242 billion), sales and excise taxes ($147 billion), state income taxes ($60 billion), and the federal individual income tax ($296 billion). Treating personal income taxes as falling fully on labor income rather than labor and capital income is something of an exaggeration, but because of the many special provisions favoring capital income contained in the income tax laws, the overstatement is probably not very large.

TABLE 1—TOTAL AND AVERAGE WELFARE COSTS, 1984 (Billions $)

	η		
m	0.2	0.3	0.4
0.38	$55.9 (7.5)	$83.8 (11.2)	$111.8 (15.0)
0.43	77.9 (10.5)	116.8 (15.7)	155.7 (20.9)
0.48	106.3 (14.3)	159.5 (21.4)	212.6 (28.5)

Note: Percentages of tax revenues that fall on labor income are shown in parentheses.

ing is the wide range of the estimates: the welfare cost when $\eta = 0.4$ and $m = 48$ percent is nearly four times as large as when $\eta = 0.2$ and $m = 38$ percent. Varying the marginal tax rate alone from 38 to 48 percent approximately doubles the total welfare cost. The wide range of estimated welfare costs that results from use of a relatively narrow range of values for the two key parameters, η and m, shows how far we are from having reliable and precise estimates of the total welfare cost. Although my preferred parameter values are 43 percent and 0.3, the available empirical evidence certainly does not rule out the other possibilities; indeed, evidence can be cited to support a higher labor supply elasticity than 0.4.

Before turning to the extension of the analysis to marginal welfare costs, two reasons why this framework may overstate the total welfare cost should be discussed. (Recall, in addition, that use of aggregate data tends to work in the opposite direction.) First, this partial-equilibrium approach assumes the marginal value product of additional hours of work is constant. With a fixed capital stock, however, an increase in labor will reduce the marginal product of labor. How large a bias is introduced by assuming a fixed wage rate depends on the elasticity of the marginal product curve relative to the labor supply elasticity. With the demand elasticity high relative to the labor supply elasticity, the degree of overstatement is small. For example, with $\eta = 0.3$ and $m = 43$ percent, assuming the marginal value product curve has an elasticity of two implies that the true welfare cost would be about 15 percent less than estimated using equation (4) and assuming the wage is constant. Moreover, the actual elasticity of the marginal value product curve is likely to be higher than two. For example, with a Cobb-Douglas technology and a labor share, α, equal to 0.75, the elasticity of the marginal product of labor curve is $1/(1-\alpha)$, or 4.0. Thus, the partial-equilibrium assumption of a fixed wage is not likely to have a quantitatively important effect on the estimation of welfare cost.[16]

The second problem is potentially more troublesome, and relates to the assumption that the compensated labor supply curve is linear which underlies the derivation of equation (4). When the supply curve is not linear, equation (4) does not provide an exact estimate of welfare cost. If, as seems likely, the actual compensated supply curve is concave, as illustrated by S in Figure 1, the estimate provided by equation (4), area ACB, will overstate the true welfare cost. The available evidence provides little basis for determining how much of a bias the assumption of linearity introduces. However, for my purposes, it is most important to note that when the approach developed here is extended to the measurement of marginal welfare cost, it is not necessary to assume linearity. Thus, estimates of marginal welfare cost may be more reliable than those of total welfare cost.

[16] Taking into account possible changes in the market wage rate when labor supply varies raises one other potentially important issue that is ignored here. When labor supply rises, the wage rate falls and the rate of return to the fixed capital stock rises. Thus, capital income rises and tax revenue from capital taxes will also rise. This general-equilibrium effect is potentially important for the estimation of marginal welfare costs that relate welfare costs to changes in revenue. Note that Stuart does not take this relationship into account in his model since he assumes that there are no taxes on capital income. It is not clear whether this effect is incorporated in the Ballard et al. (1985) model or not. Assuming a fixed wage rate, as here, sidesteps this issue since capital income is then unaffected by changes in labor supply, but the importance of this point deserves further investigation.

II. Marginal Welfare Cost

The marginal welfare cost is the ratio of the change in total welfare cost to the change in tax revenue produced when tax rates are varied in some specified way. With W representing the total welfare cost and R total tax revenue, it is simply dW/dR. Figure 2 illustrates the numerator, dW, of the marginal welfare cost ratio. When the marginal tax rate rises from m to m', there is a reduction in the quantity of labor supplied along the compensated supply curve to L_3. The increment in the total welfare cost produced by this increase in the marginal tax rate is shown by area $CDEA$.[17] Area $CDEA$ is dW; dividing this by the increase in tax revenues —which is not shown in the diagram since it does not identify what happens to either the average tax rate or the actual (as distinct from the compensated) quantity of labor—measures the marginal welfare cost of raising additional revenue from taxes falling on labor income.

An expression to estimate the marginal welfare cost can be derived easily. Note that[18]

$$(6)\qquad dW=\tfrac{1}{2}(wm+wm')\,dL_2.$$

Since m' equals $m+dm$ and dL_2 equals $[\eta L_2/(1-m)]\,dm$, (6) can be rewritten as

$$(7)\qquad dW=\left[\frac{m+0.5\,dm}{1-m}\right]\eta wL_2\,dm.$$

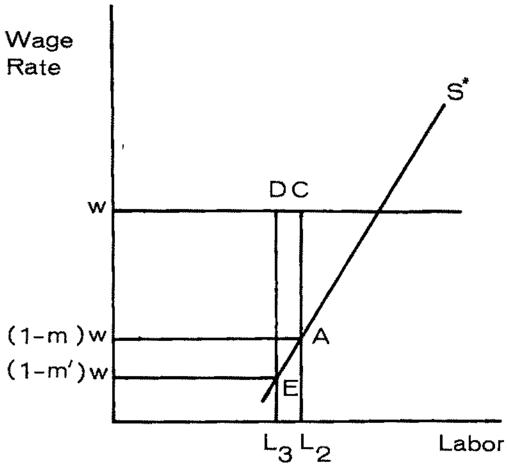


FIGURE 2

The change in tax revenue depends on how the average tax rate changes and on the change in actual labor income. It can conveniently be expressed as the sum of the additional tax revenue produced if earnings do not change and the revenue lost due to any reduction in earnings. Thus,

$$(8)\qquad dR=wL_2\,dt+w\,dL(m+dm),$$

where dt is the change in the average tax rate evaluated at the initial level of earnings, wL_2. The first term in (8) thus gives the additional revenue produced if the average rate rises by dt and labor income remains unchanged. The second term in (8) gives the revenue lost when earnings fall by wdL. Note that dL in (8) need not be equal to L_3-L_2 in Figure 2; L_3-L_2 is the compensated change in labor supply while dL is the actual change in labor supply.

Combining (7) and (8) gives us a simple expression for marginal welfare cost:

$$(9)\qquad \frac{dW}{dR}=\frac{\left[\dfrac{m+0.5\,dm}{1-m}\right]\eta wL_2\,dm}{wL_2\,dt+w\,dL(m+dm)}.$$

In principle, equation (9) can be used to evaluate marginal welfare cost for any discrete change in tax rates, but to do so requires

[17] This assumes that the incremental government expenditure restores the individual to the same indifference curve, and that the benefits from marginal government spending are a perfect substitute for disposable income, assumptions to be explained more fully later. Under these conditions, the compensated supply curve doesn't shift. Different assumptions regarding the incremental expenditures require a different interpretation of marginal welfare cost, as explained later in this section.

[18] Equation (6) depends on the assumption that the compensated supply curve is linear for the change in labor produced by the change in the marginal tax rate (dm), that is, between points E and A in Figure 2. In developing the results that follow, I assume $dm=0.01$. However, dm can be assumed to be as small as desired, and in the limit as dm approaches zero, it is, of course, not necessary to assume linearity at all.

knowledge of how actual labor earnings, the *wdL* term, will be affected. In considering the effect on actual earnings, I should begin by noting that the conceptual experiment underlying the notion of marginal welfare cost is a balanced-budget operation in which the government spends the increment in tax revenue. This implies that the marginal welfare cost of raising additional tax revenue does not depend solely on the change in the tax system, but also on how the government spends the funds.[19]

The simple theory underlying equation (9) does not take into account the full range of possible ways expenditure side effects could reinforce or offset the added tax distortions of labor supply. It can, however, take into account government expenditures in an important special case. If the marginal government spending provides benefits that are a perfect substitute for the disposable incomes of taxpayers, then the spending has only an income effect that is equivalent to a lump sum transfer. (In other words, the marginal spending can be analyzed as a parallel shift in the after-tax budget constraint.) In this case, the income effect of the spending can be taken into account through its effect on the *wdL* term in equation (9). For example, if the marginal spending, in combination with the tax change, leaves taxpayers' utilities unchanged, the actual reduction in labor earnings, *wdL*, will equal the compensated change in labor earnings and can therefore be calculated using the assumed parameter values.

Although the assumption that government spending is a perfect substitute for disposable income is restrictive, it may be more widely applicable than it first appears. Note that the marginal change in government spending does not have to take the form of cash transfers for the assumption to be valid. In particular, if the government provides a service that taxpayers would otherwise have purchased on their own, then the spending would be a perfect substitute for disposable income. This may be largely correct in cases involving government provision of schooling, medical care, pensions, and other things taxpayers would purchase with their disposable incomes if the government did not provide them. Thus, treating government expenditures as a perfect substitute for disposable income appears reasonable and permits the simple framework employed here to incorporate expenditure side effects.

Granted this assumption, there are two polar cases that seem likely to span the range of plausible outcomes. First, marginal government spending is taken to provide no benefits to taxpayers, so there is an income effect from the balanced-budget operation that acts to counter the substitution effect. I assume that the net effect on actual labor earnings is zero, so the second term in the denominator of equation (9) is zero. In this case, the formula for marginal welfare cost simplifies to

$$(10) \qquad \frac{dW}{dR} = \left[\frac{m + 0.5\,dm}{1-m}\right]\eta\frac{dm}{dt}.$$

The second polar case to be considered is when marginal government spending provides benefits that return taxpayers to their initial (i.e., before the tax and expenditure change) utility levels. When this is so, the *wdL* term in equation (9) is equal to the change in compensated labor earnings, or $-[dm/(1-m)]\eta wL_2$. Substituting this for *wdL* in (9) and simplifying yields the following expression for marginal welfare cost in this case:

$$(11) \qquad \frac{dW}{dR} = \frac{\left[\dfrac{m + 0.5\,dm}{1-m}\right]\eta\dfrac{dm}{dt}}{1-\left[\dfrac{m + dm}{1-m}\right]\eta\dfrac{dm}{dt}}.$$

Equations (10) and (11) can be used to estimate marginal welfare cost for a discrete change in marginal tax rates under the assumed conditions. This analysis indicates

[19] Several recent papers have investigated the issue of balanced-budget changes and labor supply, both from the point of view of a positive analysis of labor supply (Assar Lindbeck, 1982; James Gwartney and Richard Stroup, 1983; Arthur Snow and Ronald Warren, 1985) and in connection with the determinants of marginal welfare cost (Wildasin).

that there are four key factors that interact to determine marginal welfare cost. Two of these, η and m, were also relevant in the estimation of total welfare cost. In addition, there are two other factors that were irrelevant for total welfare costs. The first is how the balanced-budget operation affects actual labor earnings, as reflected in the wdL term in equation (9) or in the choice between equations (10) and (11) for the two special cases I will examine. Second, equations (10) and (11) show that marginal welfare cost depends also on the parameter dm/dt. This term measures the progressivity of the *change* in the tax structure that produces the incremental tax revenue. As the equations show, the more progressive the tax change (the larger dm/dt is), the greater marginal welfare cost will be.

Since there are many different ways the tax structure could be modified to produce a change in revenue, dm/dt will depend on exactly how the tax structure is changed. Thus, we must consider the range of values that dm/dt could plausibly take on. The type of change in the tax system that would probably yield the smallest value for dm/dt would be to change the rates of sales or excise taxes, or to change the Social Security payroll tax rate. Raising additional revenue by increasing the rates of these taxes implies that the marginal tax rate would rise by less than the average tax rate;[20] a reasonable assumption might be that dm/dt equals 0.8. At the other extreme, use of the federal individual income tax will typically imply that dm/dt is greater than one since this tax is progressive. With the marginal tax rate of the federal income tax nearly twice its average rate at most income levels, it seems reasonable to assume marginal tax revenue from this source implies $dm/dt = 2.0$.[21]

Between these two extremes, two other possibilities merit consideration. One is to consider a proportionate increase in the rates of all taxes simultaneously so that m/t remains unchanged. Since m equals 43 percent in my benchmark case and t equals 31 percent, this sort of change implies $dm/dt = 1.39$. The other possibility is to consider some change where dm/dt equals one; this would be appropriate if a proportional tax were added to the present tax structure. While these four values for dm/dt do not exhaust the possibilities, they probably encompass most changes we are likely to see in the tax system.

At this point, a graphical treatment of marginal welfare cost for the case in which the benefit from the expenditure returns the taxpayer to his (her) initial indifference curve may prove helpful. In Figure 3, the before-tax budget constraint relating income and leisure is YN, and the initial tax—drawn as a proportional tax for simplicity—produces the constraint Y_1N. The worker is initially at point E, with tax revenue equal to HY since HH is drawn parallel to YN. Now let us consider a small increase in the tax rate which, ignoring expenditure side effects, produces the constraint Y_2N, drawn exaggerated for clarity. Assume that the expenditure is a perfect substitute for disposable income and the benefit from the expenditure returns the

[20] An increase in the rates of sales and excise tax will reduce the real tax base of personal income taxes, and so the increment in the effective combined marginal tax rate will decline with income. To see this, suppose a general sales tax is introduced at a rate of 10 percent, and this reduces factor prices by 10 percent while the price level is unchanged. For a person in a 50 percent income tax bracket, the 50 percent rate now applies only to 90 percent of his marginal value product, so the effective marginal rate of the income tax is reduced to 45 percent, and the combined rate is 55 percent. Thus, the sales tax increased this person's effective marginal tax rate from 50 to 55 percent. By contrast, for a person initially in a 20 percent income tax bracket, the increase would be from 20 to 28 percent. For the Social Security payroll tax, the ceiling on taxable earnings implies that an increase in its rate would increase the overall average tax rate more than its weighted-average marginal tax rate.

[21] In 1984, the average tax rates at one-half median income, the median income, twice median income, and five times median income were, respectively, 5.9, 11.9, 16.0, and 26.1 percent. The corresponding marginal tax rates were 14.0, 22.0, 33.0, and 45.0 percent (Congressional Budget Office, 1984, Table VI-3). These are, however, statutory rates; the effective rates would be lower. It is also worth noting that Seater's estimate of a weighted-average marginal tax rate for the income tax in 1980 is 22.9 percent, nearly double its average rate of about 12 percent.

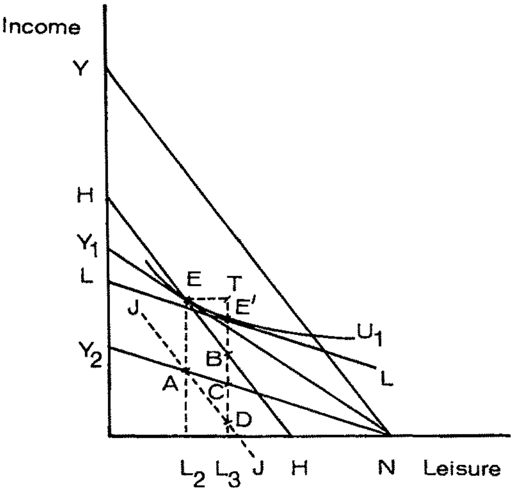


FIGURE 3

worker to his initial indifference curve. Then the effect of the expenditure can be shown as a parallel shift in Y_2N to LL, with LL tangent to U_1 at point E'.

Since the additional tax revenue is CB given the new equilibrium with labor of NL_3, the benefit from the expenditure of CB must be valued at CE' to return the worker to his initial indifference curve. Note that the required benefit, CB, is BE' greater than the additional tax revenue; BE' is the additional welfare cost. Thus, the marginal welfare cost, dW/dR, is BE'/CB. It is a compensating variation measure of the change in surplus, and shows how much greater the benefits from government spending must be than the tax revenues collected if the balanced-budget operation is to keep the worker on his initial indifference curve.

This particular way of defining marginal welfare cost produces a measure that is relevant for determining whether government expenditures combined with the taxes that finance them will leave taxpayers on balance better or worse off. In Figure 3, note that if the benefit from the expenditure of CB is anything less than CE', the worker will be worse off than he was at E, while if it is anything greater than CE', he will be better off. Put more generally, the marginal benefits from government spending must be more than one plus the marginal welfare cost (BE'/CB) per dollar spent if taxpayers are to be benefited on balance.[22] Other definitions of marginal welfare cost are possible. Stuart, for example, defines marginal welfare cost as the loss that results when the incremental tax revenue is returned to the worker as a lump sum payment. This produces a measure of the loss when outlays are valued at their budgetary cost, but it is not the appropriate definition to use in conducting a cost-benefit analysis of an expenditure policy.[23]

Note that equation (9) will estimate BE'/CB exactly. The numerator measures dW as the difference between the compensated reduction in earnings using the market wage rate, TB, less the increment in the value of leisure, TE'. The denominator

[22] Note that this measure of marginal welfare cost, based on the compensating variation, is similar to that proposed by Diamond and McFadden. The only difference is that my definition uses the utility level actually achieved with existing taxes and expenditures, whereas theirs uses the before-tax utility level. Note also that it is not inconsistent to use an equivalent variation measure of total welfare cost (as in Section I) and a compensating variation measure of marginal welfare cost. When the analysis is intended to provide a measure of marginal welfare cost useful for cost-benefit analysis, as explained in the text, the compensating variation measure is appropriate.

[23] Stuart's measure and mine yield the same result in the special case of zero income effects. In this case, if the incremental tax revenue is returned as a lump sum, the final equilibrium in Figure 3 will be at point B since an indifference curve will be tangent to the budget constraint (incorporating the lump sum transfer) that is parallel to Y_2N and passes through B. Stuart's measure is then the loss, BE', divided by the incremental tax revenue, CB. However, if leisure is a normal good, work effort will be greater than NL_3 when the tax revenue is returned as a lump sum due to the worker's loss in real income. The final equilibrium will then lie to the left of point B on the EB portion of HH, and Stuart's measure of marginal welfare cost will be smaller than BE'/CB since incremental tax revenue will be greater and the additional welfare cost will be smaller. For this case, Stuart's measure has the defect that even when the marginal expenditure is valued at one plus marginal welfare cost, the final equilibrium involves the worker being worse off than at point E because the income effect of the expenditure will lead to less work effort than the lump sum transfer. Thus, Stuart's measure does not identify how much the benefits of the expenditure must exceed additional tax revenue to exactly compensate the worker.

TABLE 2—MARGINAL WELFARE COST PER DOLLAR OF REVENUE
(Percentages)

		$m =$	0.38			0.43			0.48		
	$\frac{dm}{dt}$	$\eta =$	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
Earnings	0.8		9.9	14.9	19.9	12.2	18.3	24.4	14.9	22.4	29.8
Constant	1.0		12.4	18.6	24.8	15.3	22.9	30.5	18.7	28.0	37.3
	1.39		17.3	25.9	34.5	21.2	31.8	42.4	25.9	38.9	51.9
	2.0		24.8	37.3	49.6	30.5	45.8	61.1	37.3	56.0	74.6
Earnings	0.8		11.0	17.6	24.9	13.9	22.5	32.4	17.6	28.9	42.7
Decline	1.0		14.2	23.0	33.2	18.0	29.8	44.2	23.0	39.0	59.9
	1.39		20.9	35.1	53.1	27.0	46.9	74.3	35.1	64.1	108.9
	2.0		33.2	59.8	100.0	44.1	85.2	159.7	59.9	128.8	303.1
Average Welfare Cost			7.5	11.2	15.0	10.5	15.7	20.9	14.3	21.4	28.5

measures dR as the sum of the increment in tax revenue if earnings remain unchanged, EA ($=DB$, since JJ is parallel to HH), less the reduction in taxes due to the actual (and compensated, in this case) reduction in labor supply, CD.[24]

III. Results

To sum up, the range of values for the four key parameters that will be used here are:

m: .38, .43, and .48;
η: 0.2, 0.3, and 0.4;
dm/dt: 0.8, 1.0, 1.39, and 2.0;
wL: Unchanged, and reduced by the compensated change.

Table 2 displays the results of using equations (10) (Earnings Constant) and (11) (Earnings Decline) to calculate the marginal welfare cost for the 72 possible combinations of parameter values for an increase in the marginal tax rate on one percentage point ($dm = 0.01$). The estimates range from a low of 9.9 percent to a high exceeding 300 percent! (Note, however, that only one combination of parameter values yields an estimate exceeding 159.7 percent.) If, as I believe to be the case, our empirical evidence and theory do not allow us to narrow substantially the range of possible parameter values from those used here, then we cannot provide a very precise estimate of the marginal welfare cost. My preferred estimates are based on $\eta = 0.3$, $m = 43$ percent, and $dm/dt = 1.39$, implying that marginal welfare cost would lie between 31.8 and 46.9 percent, depending on what assumption is made about the extent to which tax payers benefit from the marginal government spending. It would be difficult, however, to defend these parameter values as necessarily more accurate than others used in the table.

The results here suggest that marginal welfare cost is significantly larger than implied by my 1976 paper. In part, the difference is due to correction of the error discussed above in Section I. The 9 to 16 percent range of my earlier paper was based on parameter values of (approximately) $\eta = 0.2$, $m = .43$, $dm/dt = 1.0$ and 1.39, with earnings constant. Table 2 shows the corrected estimates for these values would be 15.3 and 21.2 percent. The remaining difference in results, however, is due to the use here of a wider range of parameter values. What was not clear in my earlier paper, but Table 2 brings out forcefully, is how sensitive the results are to the combination of parameters used.

Even though this model is far simpler than the general-equilibrium models of Stuart and

[24] For a graphical treatment that can be used to show marginal welfare cost when the taxpayer is not returned to his original indifference curve, in which case an equivalent variation measure is used; see Figure 3 in my paper with Johnson. In this diagram, marginal welfare cost is DE/AH.

P4216

Ballard et al. (1985), the results seem quite similar for comparable parameter values. The approach used here yields estimates that are moderately larger than the Stuart model, but corrected for two differences in assumptions the results differ only negligibly.[25] Comparison with Ballard et al. is more difficult, since they are not explicit concerning all the parameter values emphasized here and their model also evaluates distortions other than the labor supply distortion. However, their general conclusion that marginal welfare cost is likely to be in the range of 15 to 50 percent accords well with the results in Table 2.

IV. Concluding Remarks

Other things the same, general-equilibrium results are to be preferred to partial-equilibrium results. Until it is shown that the general-equilibrium models provide significantly different and more accurate estimates (for the same parameter values), however, the partial-equilibrium approach has some advantages. First, it is easily understood, so it is less likely that critical assumptions will be obscured. The sensitivity of the results to the four key parameter values is quite apparent in this treatment, for example. Second, it is simple for other investigators to perform sensitivity analysis by modifying the assumptions regarding parameter values if such changes seem appropriate. Finally, on a more substantive matter, the results here seem to imply that arriving at a more precise estimate of marginal welfare cost may well depend more on empirical investigation that narrows the range of possible parameter values than on developing more rigorous models that yield slightly better estimates for given parameter values.

An important point concerning the proper use of estimates of marginal welfare cost is in order. These estimates are intended to provide the basis for comparing the costs with the benefits of government expenditure policies that do not have as a major consequence or goal a redistribution of income. Marginal welfare costs are relevant in analyzing redistributive programs, but the estimates here do not indicate how large the relevant effects are. For this purpose, it is necessary to estimate the costs borne by the group that loses separately from the benefits received by the group that gains, along the lines suggested by myself and Johnson.[26] In general, the relevant marginal welfare costs of redistribution are several times larger than the marginal welfare costs reported here. Basically, the reasons are that both the taxpayer's and recipient's decisions are distorted by a redistributive policy, and marginal tax rates necessarily rise quite sharply in comparison to the amounts redistributed.[27]

Finally, it should be recalled that the estimates relate only to the labor supply distortions of taxes. Actual taxes distort behavior on a number of other margins of choice, and ignoring these probably means that the estimates here understate the marginal welfare cost of raising tax revenue, subject to the usual second-best qualifications. Further research to incorporate these effects into the analysis would be worthwhile.

REFERENCES

Ballard, Charles L., Shoven, John B. and Whalley, John, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, March 1985, *75*, 128–38.

[25] The first difference is that Stuart defines marginal welfare cost as the loss resulting when the expenditure is a lump sum transfer back to taxpayers. The second difference is that Stuart's general-equilibrium model effectively incorporates a downward-sloping marginal value product curve. These are not the only differences in the models, but they appear to account for most of the differences in results.

[26] In this connection, it is unfortunate that Stuart refers to one of his marginal welfare cost measures as relevant for analyzing redistributional social programs. In Stuart's model, this refers to the case where the revenues are returned to the taxpayer as a lump sum payment. Since Stuart's model is based on a single aggregate household, there is no real redistribution involved in this case, and this measure of marginal welfare cost gives no clue to the relevant distortions produced by redistributive programs.

[27] I have explained in greater detail the relationship between the marginal welfare cost of raising tax revenue as discussed here and the marginal welfare cost of redistributing income in my 1986 paper (Section IV).

______, ______, and ______, "The Welfare Cost of Distortions in the United States Tax System: A General Equilibrium Approach," NBER Working Paper No. 1043, 1982.

Barro, Robert J. and Sahasakul, Chaipat, "Average Marginal Tax Rates from Social Security and the Individual Income Tax," NBER Working Paper No. 1214, 1983.

Browning, Edgar K., "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, *84*, 283–98.

______, (1985a) "The Marginal Social Security Tax on Labor," *Public Finance Quarterly*, July 1985, *13*, 227–51.

______, (1985b) "A Critical Appraisal of Hausman's Welfare Cost Estimates," *Journal of Political Economy*, October 1985, *93*, 1025–34.

______, "The Marginal Cost of Raising Tax Revenue," in Phillip Cagan, ed., *Essays in Contemporary Economic Problems*, Washington: American Enterprise Institute, 1986.

______ **and Johnson, William R.,** "The Trade-Off between Equality and Efficiency," *Journal of Political Economy*, April 1984, *92*, 175–203.

Burkhauser, Richard V. and Turner, John A., "Is the Social Security Payroll Tax a Tax?," *Public Finance Quarterly*, July 1985, *13*, 253–67.

Diamond, P. A. and McFadden, D. L., "Some Uses of the Expenditure Function in Public Finance," *Journal of Public Economics*, February 1974, *3*, 3–21.

Findlay, Christopher C. and Jones, Robert L., "The Marginal Costs of Australian Income Taxation," manuscript, Australian National University, 1981.

Gordon, Roger H., "Social Security and Labor Supply Incentives," *Contemporary Policy Issues*, April 1983, *3*, 16–22.

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–51.

Hansson, Ingemar and Stuart, Charles, "Tax Revenue and the Marginal Cost of Public Funds in Sweden," manuscript, University of California-Santa Barbara, 1983.

Harberger, Arnold C., "Taxation, Resource Allocation, and Welfare," in *The Role of Direct and Indirect Taxes in the Federal Revenue System*, NBER Other Conference Series No. 3, University Microfilms, 1964.

Hausman, Jerry A., "Labor Supply," in Henry J. Aaron and Joseph A. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: Brookings Institution, 1981.

Kay, J. A., "The Deadweight Loss from a Tax System," *Journal of Public Economics*, February 1980, *13*, 111–19.

Lindbeck, Assar, "Tax Effects Versus Budget Effects on Labor Supply," *Economic Inquiry*, October 1982, *20*, 473–89.

McGee, M. Kevin, "The Burden of Taxation Revisited," manuscript, University of Wisconsin-Oshkosh, 1985.

Seater, John J., "On the Construction of Marginal Federal Personal and Social Security Tax Rates in the U.S.," manuscript, North Carolina State University, 1984.

Snow, Arthur and Warren, Ronald S., Jr., "Labor Supply and Tax Rates in General Equilibrium," manuscript, Georgetown University, 1985.

Stuart, Charles, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, *74*, 352–62.

Wildasin, David E., "On Public Good Provision with Distortionary Taxation," *Economic Inquiry*, April 1984, *22*, 227–43.

U.S. Congress, Congressional Budget Office, *Reducing the Deficit: Spending and Revenue Options*, Washington, USGPO, February 1984.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington: USGPO, 1985.

Voluntary Donations and Public Expenditures in a Federalist System

By RICHARD STEINBERG*

Proponents of the new federalism argue that nonprofit organizations and local governments will fully offset federal social service expenditure cutbacks. I analyze this proposition as a competitive game in which donations are motivated by private and public good considerations. I characterize the response of political-economic equilibrium to exogenous changes in federal expenditures when local voters are cognizant of donor reactions. Partial replacement is the most likely outcome, though others are possible.

The nonprofit sector has long been recognized as a decentralized and pluralistic alternative to governments for the provision of such vital services as higher education, health care, research, performing arts, and public welfare (Burton Weisbrod, 1975). In 1975, the sector employed 5.7 percent of compensated labor and produced 3.2 percent of measured GNP in the United States (Gabriel Rudney, 1981). In 1982, contributions of money and volunteer time were valued at over $100 billion (Virginia Hodgkinson and Murray Weitzman, 1984). Indeed, the government is a relative newcomer in supporting many of these services.

Proponents of the new federalism view the substitution of governmental service provision for private provision by nonprofit organizations with alarm, seeking to reduce the federal role. They believe that these services are more appropriately provided by local governments and nonprofit organizations, who will automatically make up for federal cutbacks.

Previous studies have generally analyzed these propositions one piece at a time. Some studies examined the effects of federal grants on state and local expenditures (Steven Craig and Robert Inman, 1986), neglecting nonprofits. Other studies examined the effects of exogenous governmental cutbacks on donations (Burton Abrams and Mark Schmitz, 1978; 1984), neglecting all interactions between donations and endogenous local government expenditures. In this paper, I present unified models of the effect of exogenous governmental changes on aggregate donations (simple crowd out) and of the effect of federal changes on local government expenditures and donations when all interactions are taken into account. Further, my model of simple crowd out integrates the heretofore separate public and private good approaches to donations demand. This integration resolves certain counterfactual implications of the models taken separately.

I. A Model of Simple Crowd Out

A. *Model Structure*

Previous analyses of donation demand generally assumed that donors were motivated by public or by private good considerations but not both. The public good models assume that donors contribute in order to supplement provision of a public good and

*Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. Many people have made helpful comments on earlier drafts of this paper. I thank the participants of seminars at Yale, the University of Pennsylvania, V.P.I. & S.U., the University of Wyoming, and SUNY-Albany. Special thanks go to Robert Inman, Thomas Reiner, Susan Rose-Ackerman, Julian Wolpert, Avner Ben-Ner, Michael Waldman, Amoz Katz, Russell Roberts, Hervé Moulin, Burton Weisbrod, Todd Sandler, and the anonymous referees. Partial support for this research has been provided by grants from the National Science Foundation and Yale University's Program on Nonprofit Organizations.

examine properties of Nash equilibrium for this game (Martin McGuire, 1974; Weisbrod; Peter Warr, 1982; Theodore Bergstrom, Lawrence Blume, and Hal Varian, 1986; Russell Roberts, 1984).[1] However, the model has three counterfactual implications.

First, aggregate giving is likely to be very small in equilibrium due to familiar free-rider effects.[2] The current level of donations seems implausibly high compared with the predictions of this polar model.

Second, the pure public model implies dollar-for-dollar crowd out in internal (positive donation) equilibrium (Warr; Roberts, 1984).[3] This implication was rejected with statistical confidence by most empirical studies of simple crowd out. Abrams and Schmitz (1984) found that state and local social welfare payments crowded out about 30¢ per dollar (i.e., the crowd out parameter was -0.3), a value significantly different from both 0 and -1. In my 1985 paper, I found crowd out of -0.005 from British data, again significant. Karl-Heinz Paquè (1982) found crowd out between -0.06 and -0.35 for German (FRG) government social service spending, again significant.[4]

Finally, the pure public model implies that whenever government spending is positive in political equilibrium, donations will be zero (Roberts, 1984; 1985).[5] This is a direct implication of dollar-for-dollar crowd out, for marginal governmental expenditures would not supplement private giving unless donations were driven to zero by inframarginal spending. Since there are a variety of goods supported both by governments and donations, the model is inconsistent with the real world.

The private good approach stresses the personal rewards to giving. Sometimes the reward is explicit, as when the donor receives a nice newsletter about his former classmates, front row seats, public recognition, or the ability to redirect the composition of output for personal gain (Gordon Tullock, 1971). Sometimes the reward is less tangible—as giving assuages social pressures from family, friends, and employers (Steven Long, 1976). Sometimes, the reward comes from the act of giving itself (Kenneth Arrow, 1974), as the donor acquires the feeling that she has satisfied her Kantian categorical imperative (Thomas Ireland and David Johnson, 1970), contributed her fair share (Susan Rose-Ackerman, 1982), or received some related feeling of satisfaction (Howard Margolis, 1981; Robert Sugden, 1984).

The purely private goods model has the counterfactual implication that crowd out should be zero, which is inconsistent with the previously cited studies of crowd out. While the private goods model can be rescued in a variety of ways,[6] the method employed here is to combine it with a public

[1] Two exceptions are Jeffrey Weiss (1981), who looks for core solutions in a cooperative game incorporating government endogenously, and Bruce Seaman (1979), who looks at myopic Cournot-type dynamics rather than equilibrium.

[2] For a nice exposition of this result, see Anthony Atkinson and Joseph Stiglitz (1980, pp. 505–07).

[3] Although both authors limit attention to the case of income redistribution, their 100 percent crowd out conclusion logically extends to all pure public goods and to all publicly provided goods with fixed distribution shares. The conclusion depends on two assumptions—that expenditures by any other party are a perfect substitute for one's own donation, and that the set of agents making positive donations is fixed. Bergstrom et al. relax the second assumption and find some exceptions to the total crowd out conclusions, but B. Douglas Bernheim (1986) demonstrates that when there is sufficient overlap between the sets of donors to different causes, the perfect-substitute assumption is even less plausible. Not only is there 100 percent crowd out, but all taxes are equivalent to lump sum taxes! In this paper, I remedy these counterfactual implications by relaxing the first assumption.

[4] Other empirical studies of crowd out include Abrams and Schmitz (1978), William Reece (1979), Orley Amos (1982), myself (1983), Philip Jones (1983), and Jerald Schiff (1985).

[5] Roberts (1985) points out one exception to this conclusion. When a physical good is provided publicly and some donors receive rents from the production of this good, a sufficiently large income effect could induce these donors to continue positive giving in political equilibrium. However, he shows that giving would be zero for all other donors in political equilibrium of a pure public model, which seems counterfactual.

[6] Rose-Ackerman (1981) details five potential sources of positive and negative crowd out in a private goods model—economies of scale, asymmetric information, endogenous ideologies, complementarities, and matching grants. These sources are neglected in this paper to focus on other issues.

goods model. I assume that the act of giving provides additional utility which is distinct from the utility obtained from the aggregate level of provision of the public good. Thus, I assume preferences are representable by the following twice differentiable and strictly quasi-concave utility function:

$$U_i = U_i(C_i, D_i, \bar{P}_i), \tag{1}$$

where C_i is consumption of a composite private good by person i, D_i is i's voluntary donation, $\bar{P}_i$ is the total level of the public good provided by others ($\bar{P}_i = \Sigma D_{j_{j \neq i}} + G^L + G^F$), G^L is local government expenditure of the public good, and G^F is federal expenditure on the (local) public good in the community.

Presumably, D_i and $\bar{P}_i$ are Hicksian substitutes for most individuals, as the joy of giving would be greater when others fail to meet "charitable needs." However, the two might be complements when giving is motivated by a comparison with other members of some reference group. This specification is consistent with either relation. The special case where D_i and $\bar{P}_i$ are perfect substitutes corresponds to the pure public goods case. The private goods case occurs when $\bar{P}_i$ has a vanishing impact on i's utility.[7]

The typical donor sees three sectors supporting provision of some good—the federal government, his local government, and the nonprofit sector. I assume that this donor is indifferent between provision by other donors, the federal government, and his local government. Thus, only the sum of expenditures by these groups centers the donor utility function.[8] The key defining characteristic of nonprofit organizations in this model is that they receive substantial resources from donations. All fund-raising expenditures and other sources of nonprofit receipts are assumed exogenous.[9]

I specify utility in terms of expenditures because I assume that marginal cost is constant and identical across sectors.[10] This good has local publicness in that expenditures on the good enter each individual's utility function in a particular governmental jurisdiction. This could be because the good is public in the Samuelsonian sense, or because the good is private in nature, but is publicly provided to each member of the community according to a fixed allocation rule. For simplicity, I assume that the good cannot be purchased on an individual basis.[11]

[7] It might seem that the private goods case corresponds to unrelatedness of D_i and P_i, but this restriction is insufficient to eliminate the possibility of crowd out. As I demonstrate below, quantity rations can have income effects (of either sign) on consumption of each other good even when the other good is unrelated in the Hicksian sense.

[8] Working independently, Schiff adopted a similar specification of simple crowd out and obtained similar results. In his model, aggregate government spending and the giving of others are imperfect substitutes for each other as well as for own donations. This generalization has two costs—he was forced to assume that donors are identical and to use a non-Hicksian definition of imperfect substitutability. In light of these complications and the seeming robustness of comparative static results, I feel comfortable making the more restrictive but simpler assumption that local government spending, federal spending, and the giving of others are perfect substitutes.

[9] This assumption is far from innocuous, but allows us to focus more clearly on certain issues. See my paper (1986) and Rose-Ackerman (1985) for models of the effect of federal grants to nonprofit organizations with endogenous fundraising expenditures. To my knowledge, no one has yet analyzed the impact of direct governmental spending in a model with endogenous fund-raising expenditures.

[10] The assumption that marginal cost is identical across sectors is inessential. All that is required is fixed relative prices—then changes in expenditure will mirror changes in quantity.

[11] This assumption may not be innocuous. Many goods provided by nonprofit firms (such as health care, day care, or nursing homes) are also provided by for-profit firms. However, the goods provided by donations to nonprofit firms are typically very different from the goods purchased from either nonprofit or for-profit firms. Thus, medical care is purchased from either for-profit, nonprofit, or governmental hospitals, but donations to nonprofit hospitals "purchase" research, capital improvements, or charity care. Therefore, it may be appropriate to lump all purchases together with the composite private consumption good, though this needs to be rationalized in an explicit model. Other goods (such as welfare or disaster relief) are provided only by governments and nonprofit organizations. The model presented here is best suited for these goods.

B. *The Individual Choice Problem*

Local taxes are approximated by a proportional income tax which fully finances local governmental expenditure on the public good. Thus:

$$T_i^{L,h} = Y_i G^{L,h}/Y^h \tag{2}$$

where $T_i^{L,h}$ is the local taxes of the ith taxpayer in the hth community, Y_i is i's income, and Y^h is aggregate income in community h.

Federal taxes are assumed linear and progressive on taxable income. Local taxes and charitable donations are deductible, as everyone itemizes.[12] Thus (leaving superscript h understood):

$$T_i^F = \alpha + \beta\left[Y_i\left(1-\left(G^L/Y\right)\right) - D_i\right]; \qquad \alpha < 0, \quad \beta > 0. \tag{3}$$

Donor i will pick C_i and D_i to maximize utility subject to a budget constraint and a quantity constraint which reflects expenditures on the good by government and others. The Lagrangian for this problem is

$$\begin{aligned} L_i = U_i(C_i, D_i, \overline{P}_i) &+ \lambda_1\big(Z_i - p_c C - (1-\beta)D_i \\ &- (1-\beta)\left(Y_i G^L/Y\right)\big) + \lambda_2\left(\overline{P}_i - a_i\right); \\ &(C_i, D_i) \geq 0, \end{aligned} \tag{4}$$

where $a_i \equiv G^L + G^F + \overline{D}_i$, and represents the ration level (outside i's control) of the public good, $\overline{D}_i = \Sigma_{j \neq i} D_j$ (total donations of others in the community, $Z_i = Y_i(1-\beta) - \alpha$ (income after federal taxes in the absence of itemized deductions or exogenous income), and p_c is the price of private consumption, C relative to the (numerairre) marginal cost of the public good.

This problem can be solved to produce a set of $2N$ individual ordinary (uncompensated) conditional demand functions of the form:

$$C_i^* = f\left(Z_i, p_c, 1-\beta, (1-\beta)Y_i/Y; a_i\right); \quad i = 1, \ldots, N, \tag{5a}$$

$$D_i^* = g\left(Z_i, p_c, 1-\beta, (1-\beta)Y_i/Y; a_i\right); \quad i = 1, \ldots, N. \tag{5b}$$

Thus, donation demand depends upon the exogenous component of income, the relative tax prices of consumption, donations, and local government spending, and the ration level of the public good.

C. *Simple Crowd Out*

The comparative statics of individual donation demand are standard except for the effect of a change in the quantity constraint a_i. James Tobin and Henrik Houthakker (1951) derived general comparative static results for variation of the quantity constraint in a neighborhood around the unconstrained equilibrium, but these well-known results are not quite appropriate here. In a community of diverse preferences with a single level of public good provision and non-Lindahl taxes, very few donors will find themselves in a neighborhood around their unconstrained optimum. More recently, Robert Mackay and Gerald Whitney (1980) derived a Slutsky-equation analog for variations in quantity constraints in any neighborhood. In the present notation, they showed:

$$\frac{\partial D_i^*}{\partial a_i} = \frac{\partial D^U}{d a_i} + \frac{\lambda_{2i}}{\lambda_{1i}} \frac{\partial D_i^*}{\partial Z_i}, \tag{6}$$

where superscript U indicates a compensated (utility held constant) derivative.

The first term is a substitution effect which has the same sign as the Hicksian substitu-

[12] Most taxpapers do not itemize. Fortunately, the assumption of the text is stronger than necessary. For nonitemizers, the relative price of local government spending to donations remains the same. I need only rule out endogenous itemization effects. For complications resulting when itemization status is endogenous, see Reece and Kimberly Zieschang (1985).

tion effect. Thus, if donations and public goods are (compensated) substitutes, this term is negative. The second term represents an income effect. Unlike the unconstrained case, the income effect involves λ_{2i}, the Lagrange multiplier on the quantity constraint. Typically, some donors will feel undersupplied and others will feel oversupplied in the constrained good (given their coerced tax payments). For undersupplied donors, a relaxation of the quantity constraint acts like an increase in income, so that nonsatiated individuals will, *ceteris peribus*, increase their donations when donations are a normal good. Similarly, relaxation of the quantity constraint acts like a decrease in income for oversupplied donors.

In summary, the Mackay-Whitney decomposition applied to nonsatiated individuals implies that when D_i and $\bar{P}_i$ are Hicksian substitutes and $\bar{P}_i$ is normal:

$\partial D_i^*/\partial a_i < 0$ if $\bar{P}_i \geq \bar{P}_i^*$;

sign $\partial D_i^*/\partial a_i$ is ambiguous otherwise.

When D_i and $\bar{P}_i$ are Hicksian complements:

$\partial D_i^*/\partial a_i > 0$ if $\bar{P}_i \leq \bar{P}_i^*$;

sign $\partial D_i^*/\partial a_i$ is ambiguous otherwise.

Thus, when the public good is a substitute for one's own donations, the derivative can only be signed if the public good is optimally or oversupplied. When the public good is a complement to own donations, the sign is only definite if the public good is optimally or undersupplied.[13]

The innovation of this section is to show that this decomposition applies to charitable giving, and to analyze how it aggregates when the quantity constraint is partly the outcome of a Nash donations game. I present sufficient conditions for existence and uniqueness of equilibrium in an unpublished appendix (available upon request) and these conditions do not appear overly restrictive.

In order to derive the desired result, I first relate changes in individual conditional donation demand curves to changes in government spending. I assume that initial and final donations are nonzero for tractability (but see fn. 3). Local government spending enters (5b) only through its effect on the quantity constraint, thus:

$$(7)\quad \left.\frac{dD_i^*}{dG^L}\right|_{BB} = \frac{\partial D_i^*}{\partial a_i}\left.\frac{da_i}{dG^L}\right|_{BB} = \frac{\partial D_i^*}{\partial a_i}\left(1+\left.\frac{d\bar{D}_i}{dG^L}\right|_{BB}\right),$$

where $|_{BB}$ indicates the total derivative encompassing a balanced budget change in local government spending, with federal spending, income, and relative prices held constant.

Equation (7) is the derivative of i's best-response function. Since the derivative of the best-response function depends only on the sum of donative strategies of other plays (not the whole vector), a fixed point of the system (which is Nash in changes) is found by aggregating (7) across i and solving for the aggregate donative response. Thus:

$$(8)\quad \left.\frac{dD^*}{dG^L}\right|_{BB} = \frac{\partial D^*}{\partial a} + \sum_i \left.\frac{d\bar{D}_i}{dG^L}\right|_{BB}\frac{\partial D_i^*}{\partial a_i},$$

[13] To determine whether the public good is oversupplied, one must calculate its implicit price. When the public good is supplied through a balanced budget change in G^L, the price i must pay depends on both the tax rule and the donative reaction to increased G^L. It is easy to show that

$$p_{G^L}|_{BB} = ((1-\beta)Y_i/Y)/(1+(dD/dG^L)/_{BB}).$$

Thus, whether a particular donor feels the public good is undersupplied depends on his or her preferences, relative income, and the crowd out parameter evaluated at current exogenous $G^L + G^F$. This result should be contrasted with that of Richard Cornes and Todd Sandler (1984). Their model is analogous to the present one, though it is developed somewhat differently. They examine the case in which a purchase is jointly an attribute of a private good and a public good. They show that the sign of crowd out is ambiguous in the complements case, but do not show sign definiteness for complements or ambiguity for substitutes when the public good is oversupplied. The reason is that they do not consider any cases in which donors would feel over supplied, for they examine an equilibrium with Nash conjectures and no government. In this situation, the effective price of giving is zero so that nonsatiated individuals would always feel undersupplied in the public good.

where

$$\left.\frac{dD^*}{dG^L}\right|_{BB}=\sum_i\left.\frac{dD_i^*}{dG^L}\right|_{BB}\quad\text{and}\quad\frac{\partial D^*}{\partial a}=\sum_i\frac{\partial D_i^*}{\partial a_i}.$$

Although an exact solution to (8) is available,[14] an approximate solution is more enlightening. Assume that

$$\left.\frac{d\bar{D}_i}{dG^L}\right|_{BB}=\left.\frac{d\bar{D}_j}{dG^L}\right|_{BB}\equiv\left.\frac{dD^*}{dG^L}\right|_{BB}\quad\text{for all } i, j,$$

which is approximately true when i contributes a small share of total donations. Then,

$$(9)\quad \left.\frac{dD^*}{dG^L}\right|_{BB}=\frac{\partial D^*}{\partial a}\Big/\left(1-\frac{\partial D^*}{\partial a}\right).$$

Now partition the community into two sets of individuals—the nonundersupplied (set J) and the undersupplied (set K):[15]

$$J=\{i:\bar{P}_i^*\leq\bar{P}_i\};\quad K=\{i:\bar{P}_i^*>\bar{P}_i\}$$

Substituting the Mackay/Whitney decomposition into (9) for each set, one obtains:

$$(10)\quad dD^*/dG^L|_{BB}=(\theta-\psi)/(1+\psi-\theta),$$

$$\text{where}\quad \theta=\frac{\lambda_{2K}}{\lambda_{1k}}\frac{\partial D_K^*}{\partial Z_K};\quad \theta>0$$

$$\text{and}\quad \psi=-\frac{\partial D^V}{\partial a}-\frac{\lambda_{2J}}{\lambda_{1J}}\frac{\partial D_J^*}{\partial Z_J};\quad \psi>0.$$

That is, θ represents the aggregate of perverse income effects, ψ represents all other effects.

There are four cases to consider. In the first case, the substitution effect overwhelms the income effect for donors in K. In this case, both summations are negative, so that total donations in a community rise by a positive fraction of the decrease in government spending. Crowd out is partial because the numerator is opposite in sign and smaller in absolute value than the denominator. In the second case, the summation over donors in K is positive, but smaller in absolute value than the (necessarily negative) summation over J. Here, too, one can verify easily that crowd out is incomplete.

However, when the summation over K is positive, greater in absolute value than over J, but less than this absolute value plus one, negative crowd out occurs. Finally, if the summation over K is greater than this, the sign of dD^*/dG^L is once again negative but the magnitude of this derivative is smaller than negative 1, a phenomenon denoted super crowd out.

The negative and super crowd out cases seem unlikely, as they require the income effect to dominate the substitution effect for undersupplied donors. In addition, they require that the demands of the undersupplied dominate the aggregate donations function. Although most empirical studies find that crowd out is partial, Paquè found a positive elasticity of giving (hence negative crowd out) for German government spending on "health and recreation" and on "cultural affairs," with the latter significantly positive. Thus, the model developed here is consistent

[14] To find an exact solution, substitute

$$\left.\frac{dD}{dG^L}\right|_{BB}-\left.\frac{dD_i^*}{dG^L}\right|_{BB}\quad\text{for}\quad\left.\frac{d\bar{D}_i}{dG^L}\right|_{BB}$$

in equation (8), solve for $dD_i^*/dG^L|_{BB}$, then aggregate and solve again to obtain:

$$dD/dG^L|_{BB}=k/(1-k)$$

$$\text{where}\quad k=\sum_i\left[\frac{\partial D_i^*}{\partial a_i}\Big/\left(1+\frac{\partial D_i^*}{\partial a_i}\right)\right]$$

The Mackay-Whitney decomposition applied to this exactly aggregated equation yields results similar to those in the text.

[15] Since perverse results stem only from the K set, one must show that it is nonempty for plausible situations. Since the denominator of price in fn. 13 is identical for all i, price varies with relative income. In a world where all i have identical preferences but income varies, lower-income individuals will wish to see a greater provision level than higher-income individuals, and political outcomes are therefore likely to leave at least one individual undersupplied.

both with the bulk of empirical studies and with the occasional outliers.

II. A Model of Joint Crowd Out

Local government spending is sensitive to donations and vice versa. In this section I characterize political-economic equilibrium and demonstrate how it is perturbed by exogenous federal grants.

I assume that voter preferences are single-peaked over aggregate provision of the public good,[16] and that the policy space is one-dimensional. Thus, a unique political equilibrium exists at the most preferred point of the median voter (Duncan Black, 1958).[17] The median voter picks a level of government spending (and a corresponding tax rate which balances the local budget) so as to maximize his utility, which depends on G^L because of its effect on total P. Further, I assume that the effect of G^L on D is known and utilized by voters.

Donating is a game while voting is not. The optimal donation for i depends on donations by $j \neq i$, which is not known in advance. In contrast, the function relating aggregate donations to government spending is well defined and known in advance, so that the voter problem has a well-defined maximum.

The Lagrangian for the decisive voter is

$$(11)\quad L = U(C_i, D_i, \pi(G^L)) + \lambda\big[(1-\beta)Y_i - (1-\beta)(Y_i/Y)G^L - p_c C_i - \alpha - (1-\beta)D_i\big]$$

where $\pi = \bar{D}_i(G^L|G^F$, local budget balance$) + G^L + G^F$. By manipulating the first-order conditions, I obtain

$$(12a)\quad MRS_{CD} = p_c/(1-\beta),$$

$$(12b)\quad MRS_{CP} = (p_c Y(1+D_G))/((1-\beta)Y_i)$$

$$(12c)\quad MRS_{DP} = (Y/Y_i)(1+D_G)$$

where MRS_{ab} is the marginal rate of substitution between a and b and $D_G = d\bar{D}_i/dG^L|_{BB}$.

These first-order conditions have the usual interpretation—marginal rates of substitution should be equated with relative price ratios. The relative price of C to D is $p_c/(1-\beta)$ because donations are deductible. Both local government spending and donations are deductible, so the marginal federal tax rate drops out of (12c), though the financing of local spending introduces a term reflecting relative income—Y/Y_i.

The interesting price component is $(1+D_G)$, where D_G is the derivative of donations by others with respect to financed local government spending. This is like a price because it indicates how much of the utility argument good (P) is purchased by a unit of your instrumental good (G^L). If D_G is close to (but greater than) negative one, then the price of the public good when purchased through increased G^L is high, for it requires a large increase in spending on G^L to accomplish a unit increase in P. Thus, changes in D_G will have substitution and income effects.

In turn, the properties of π, especially the sign and value of D_G, result from the short-run conditional optimization of donors described in Section I. This is because following an election, the new level of G^L is once again a quantity constraint on donations.

[16] Since voters vote only for *public* provision of the public good, single-peakedness in aggregate provision is insufficient to assure existence of a unique median-voter equilibrium. I therefore make the additional assumption that the aggregate donations function is monotone with respect to balanced-budget local government expenditure increases, which is sufficient for the desired result.

[17] Ireland-Johnson, Weisbrod, and Julian Wolpert (1977) each make similar assumptions in informal models. They argue that when public provision is governed by the preferences of the median voter, one-half of the population will be dissatisfied and wish to supplement public provision with their private donations. The informal setting is not well suited for examining feedback effects, as it doesn't allow analysis of the response of the median voter to the prospect of private supplementation. Seaman allowed a simple feedback from donations to public provision, but he examined only short-run effects rather than properties of Nash equilibria. Weiss examined the core of a cooperative voting/donor game. Roberts assumed that politicians set public expenditure levels to maximize their share of political support from voter/donors.

Partial crowd out is again most likely, implying D_G is between zero and negative one.

The other properties of political-economic equilibrium are easily derived graphically when I make two simplifying assumptions. First, I assume that the decisive voter is not a donor. Since each donor's contributions constitute a negligible portion of total donations, this assumption seems innocuous.

Second, I assume that the marginal response of donations to local government spending is the same as the response to federal grants, *ceteris paribus*, even though the former is inherently accompanied by a tax change while the latter may not be. (The grant may be a retargetting toward a community rather than an increase in federal spending.) This allows us to write D as a function of $G^L + G^F$, with derivative D_G. It seems unlikely that qualitative conclusions would be affected by this simplification.

Under these assumptions, five conclusions are (demonstrated in the Appendix):

1) If D_G is less than or equal to negative one on the margin, then local government spending will be zero in political equilibrium *regardless* of the preferences of the decisive voter. Thus, marginal super (simple) crowd out can only occur when the political system is not in equilibrium.

2) If D_G is close to (but greater than) negative one, the voter's choice set may be nonconvex, leading to the usual uniqueness and continuity of equilibrium problems.

3) A federal grant will generally cause total spending on the public good (federal and local government and local donations) to rise, but by less than the grant. This will be denoted "partial marginal joint crowd out." Sufficient conditions for partial joint crowd out are that both goods are normal, the median voter's choice set is convex, political-economic equilibrium is interior, and donations are locally linear with D_G negative.

4) When D_G is positive, partial joint crowd out is likely, but zero or negative (total spending rises by more than the grant) joint crowd out are possible.

5) Local government spending from its own tax base will fall in response to a federal grant (though local government spending, including the grant, will rise). Donations may rise or fall in political-economic equilibrium regardless of the sign of D_G.

In principle, these propositions could be verified empirically. In practice such tests would be quite difficult, for there are very little data currently available on giving at the local level.

Simple models of intergovernmental grants predict that a lump sum federal grant should have the same effect on local government spending (including the grant) as an increase in community income (in interior equilibrium). Yet, study after study has found that federal grants have bigger impacts than income increments of the same size. This so-called flypaper effect (money sticks where it hits) has inspired recent explanations involving nonlinearities in the decisive voter's budget set (Robert Moffitt, 1984), the technology of public good production (Bruce Hamilton, 1983), or games between bureaucrats or representatives and voters (Thomas Romer and Howard Rosenthal, 1979; Radu Filimon, Romer, and Rosenthal, 1982; and Craig-Inman).

The joint crowd out model suggests that there is even greater need to explain the flypaper effect. An increase in the money income of each citizen in a community will cause donations to rise (assuming they are normal). Thus, the real income of the decisive voter goes up by more than his nominal income. This increment to real income is given by

$$dY_i^R/dY_i^N = 1 + D_Y, \tag{13}$$

where superscript R indicates real value, N indicates nominal value. D_Y is the derivative of the aggregate donations function with respect to community income.

A federal grant will affect donations differently, but once again the increment to the real income of the decisive voter will not generally equal the amount of the grant. In interior equilibrium, this increment is given by

$$dY_i^R/dG^F = 1 + D_G. \tag{14}$$

Rearranging (13) and (14) and noting that

$$\frac{dG^*}{dG^F}=\frac{dG^*}{dY_i^R}\frac{dY_i^R}{dG^F} \quad \text{and} \quad \frac{dG^*}{dY_i^R}=\frac{dG^*}{dY_i^N}\frac{dY_i^N}{dY_i^R}$$

where $G^*=G^{L*}+G^F$, I obtain

$$\frac{dG^*}{dG^F}=\frac{dG^*}{dY_i^N}\frac{(1+D_G)}{(1+D_Y)}$$

Thus, a federal grant should only have the same effect as an increase in community income if $D_G=D_Y$. When donations are a normal good and crowd out is partial, these two terms will have opposite signs as well as unequal magnitude, and the model predicts that a federal grant should have a smaller effect than an equal increment to nominal community income. My model, as it stands, is inconsistent with existing empirical results on the flypaper effect. On the other hand, my model suggests that existing tests of whether the flypaper effect can be explained by nonlinearities, technology, or politics are misspecified. Future researchers of intergovernmental grants should consider incorporating joint crowd out effects in their studies.

III. Summary

One of the main tenets of the new federalism is that local governments and nonprofit organizations will increase their expenditures on social services in response to federal cutbacks. Here, I examine that proposition with a model in which donors receive private benefits which are distinct from (but related to) the benefits provided by incremental service provision. This model remedies certain counterfactual implications of previous models of giving by integrating the private and public goods approaches. I show that the sign and magnitude of the donative response to exogenous changes in government spending are each ambiguous, depending on whether donations are normal or inferior, complementary or substitutable for public expenditures, and on whether provision of the public good by others is higher or lower than the donor's unconstrained optimum. It seems most likely that donations will make up for only a portion of governmental cutbacks (denoted partial simple crowd out), though it is possible that donations would rise by more than the cutback or that donations would fall.

However, government spending is not exogenous. In Section II, I examined the joint impact of exogenous federal cutbacks on enogenous local government spending and local donations. Voters take both the level and the reaction of donations to government expenditure changes into account when choosing a level of local government spending. In communities with near total donative crowd out, the price of local government spending becomes very high and the government share of total expenditures becomes correspondingly lower. Under certain circumstances, the voter choice set will not be convex, leading to the usual problems of multiple equilibria, discontinuity of equilibria, and likely corner solutions. However, when the choice set is convex, partial joint crowd out is likely regardless of the sign of the simple crowd out parameter. That is, whether donations rise or fall in response to an exogenous federal cutback, it is likely that the total of donations and local government expenditure will rise, but only by some fraction of the cutback. One should not count on the local and private sectors to replace the federal government's role in social service provision.

Appendix

Graphical Analysis of Political-Economic Equilibrium

To begin, assume $G^F=0$ and, without confusion, denote G^L by G. The median voter has preferences over P and C, to be optimized by choosing G and C. His (her) budget constraint can be expressed as

$$\text{(A1)} \quad P(C)=Y\left(1-\frac{\alpha}{(1-\beta)Y_i}-\frac{p_cC}{(1-\beta)Y_i}\right)$$
$$+D\left[Y\left(1-\frac{\alpha}{(1-\beta)Y_i}-\frac{p_cC}{(1-\beta)Y_i}\right)\right]$$

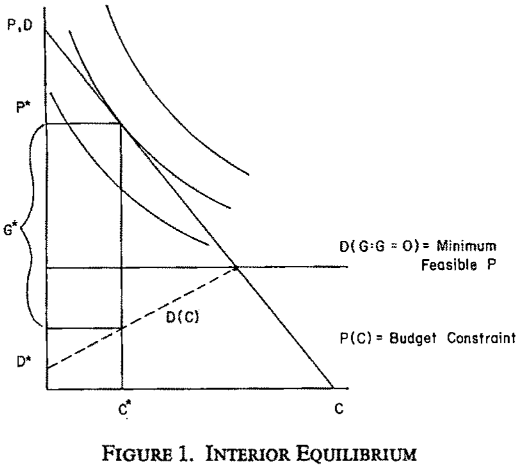


FIGURE 1. INTERIOR EQUILIBRIUM

where square brackets indicate functional dependence rather than multiplication. Since D remains positive for $G=0$, the feasible set is bounded by (A1) and a horizontal line whose P coordinate is $D(0)$. If this set is closed and convex, necessary and sufficient conditions for optimality of interior solutions are provided by the usual unique tangency of the budget constraint and indifference curves, as illustrated by Figure 1. However, the reaction of donors to government spending makes convexity of the feasible set problematic, while the nonattainability of certain positive values of the public good makes interiority problematic.

Convexity hinges on the derivatives of the donative function. Below, I will generally assume that the compensated donations function is linear in the quantity constraint and that the ordinary donations function is linear in income, which is sufficient to insure convexity (except for a possible kink, discussed below). If $D_G > -1$, then the budget constraint is a downward-sloping straight line so that convexity is no problem. If, instead, $D_G \leq -1$, the budget constraint would be upward sloping. Here, the choice set is bounded instead by the nonnegativity of G. The $G=0$ corner is optimal regardless of the decisive voter's preferences. Returning briefly to the nonconstant D_G case, when $D_G \leq -1$ for only some feasible C, one finds an interior solution where $D_G > -1$, since tangency with the voter's indifference curves can only occur on downward-sloping sections of the budget constraint. Thus, although D_G need not exceed negative one everywhere, it must exceed this value when political equilibrium involves positive public spending, establishing the first proposition of the text.

Assuming D_G is greater than negative one, a simple transformation allows one to identify changes in G directly from a diagram in (P,C) space. Utilizing the ordinary budget constraint, one defines $G=f(C)$. Thus, $D(G)$ can be written as $D(f(C))$, or simply $D(C)$. The upward-sloping straight line segment $D(C)$ meets the intersection of $D(G\colon G=0)$ (or simply $D(0)$) and the budget constraint $P(C)$. Then $G(C)$ can be found as the vertical difference between the budget constraint $P(C)$ and $D(C)$, as in Figure 1.

Because D is nonnegative, $D(C)$ may be zero for a range of values of C. This is more likely if D_G is close to negative one, for then decreases in C lead to increases in G which may rapidly push D to zero. If $D=0$ for some feasible G, then the $P(C)$ budget constraint is kinked above that point. To the right of that point, marginal sacrifices of consumption purchase less of the public good because donations fall, but this does not happen to the left of the point. This too renders the feasible set nonconvex, establishing the second proposition of the text. Subsequent results assume either that this kink does not occur or that the local optimum with positive donations dominates the zero donation optimum.

Now, consider the result of increasing federal grants from zero to some positive level when D_G is negative and greater than negative one. In this more common case, $D(0)$ (donations in the absence of local government spending) will fall by less than the grant, while the level of public goods in the absence of local government spending (denoted $D(0)+G^F$) will rise. $D(0)+G^F$ constitutes a new lower limit on attainable P, and the voters new $P(C)$ budget line rises from this point with the same slope as the old one (again, this assumes D_G is constant).

The situation is illustrated in Figure 2. Since both goods are normal and the grant acts like an increase in income when equi-

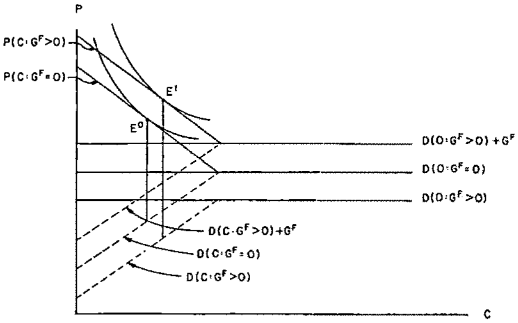


FIGURE 2. EFFECT OF A FEDERAL GRANT WHEN $D_G < 0$

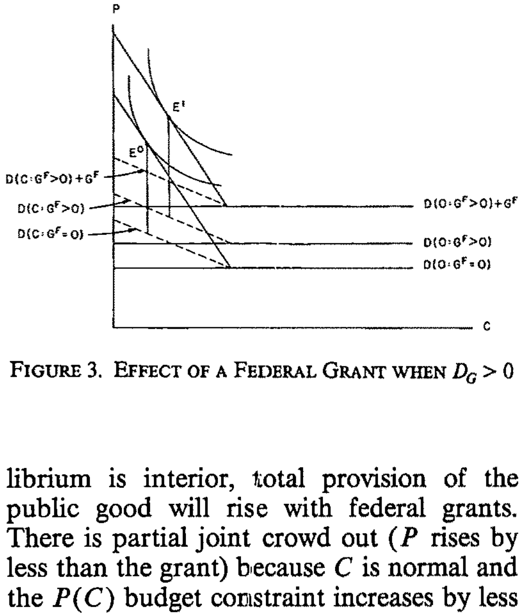


FIGURE 3. EFFECT OF A FEDERAL GRANT WHEN $D_G > 0$

librium is interior, total provision of the public good will rise with federal grants. There is partial joint crowd out (P rises by less than the grant) because C is normal and the $P(C)$ budget constraint increases by less than the federal grant. Local government spending after the grant is given by the vertical distance between E^1 and $D(C)+G^F$. As long as C is normal, it is easy to show (by similar triangles) that G^L falls in response to the grant. Finally, local donations may rise or fall. If G^L falls by more than G^F rises, D will increase. As long as the crowd out of G^L by G^F is partial, crowd out of D by G^F will be, too.

In the rarer case where D_G is positive, total spending on the public good may rise by less, as much, or more than the federal grant (thus joint crowd out is partial, zero, or negative). Figure 3 illustrates this conclusion, again, assuming both goods are normal. One can show (by similar triangles) that G^L must fall, but one cannot make any general prediction about whether donations will be higher or lower in the new political-economic equilibrium. These results establish the third, fourth, and fifth propositions of the text.

REFERENCES

Abrams, Burton and Schmitz, Mark D., "The 'Crowding-Out' Effect of Governmental Transfers on Private Charitable Contributions," *Public Choice*, No. 1, 1978, *33*, 29–37.

______ and ______, "The Crowding-Out Effect of Governmental Transfers on Private Charitable Contributions: Cross-Section Evidence," *National Tax Journal*, December 1984, *37*, 563–68.

Amos, Orley M., Jr., "Empirical Analysis of Motives Underlying Individual Contributions to Charity," *Atlantic Economic Journal*, December 1982, *10*, 45–52.

Arrow, Kenneth J., "Gifts and Exchanges," *Philosophy and Public Affairs*, Summer 1974, *1*, 343–62.

Atkinson, Anthony and Stiglitz, Joseph, *Lectures on Public Economics*, New York: McGraw-Hill, 1980.

Bertstrom, Theodore, Blume, Lawrence and Varian, Hal, "On the Private Provision of Public Goods," *Journal of Public Economics*, June 1986, *29*, 25–49.

Bernheim, B. Douglas, "On the Voluntary and Involuntary Provision of Public Goods," *American Economic Review*, September 1986, *76*, 789–93.

Black, Duncan, *The Theory of Committees and Elections*, New York: Cambridge University Press, 1958.

Cornes, Richard and Sandler, Todd, "Easy Riders, Joint Production, and Public Goods," *Economic Journal*, September 1984, *94*, 580–98.

Craig, Steven G. and Inman, Robert P., "Education, Welfare, and the New Federalism: State Budgeting in a Federalist Public Economy," in Harvey Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1986.

Filimon, Radu, Romer, Thomas and Rosenthal, Howard, "Asymmetric Information and

Agenda Control: The Basis of Monopoly Power in Public Spending, *Journal of Public Economics*, February 1982, *17*, 51–70.

Hamilton, Bruce, "The Flypaper Effect and other Anomalies," *Journal of Public Economics*, December 1983, *22*, 347–62.

Hodgkinson, Virginia and Weitzman, Murray, *Dimensions of the Independent Sector: A Statistical Profile*, Washington: Independent Sector, 1984.

Ireland, Thomas R. and Johnson, David, *The Economics of Charity*, Blacksburg: Center for the Study of Public Choice, 1970.

Jones, Philip R., "Aid to Charities," *International Journal of Social Economics*, No. 2, 1983, *10*, 3–11.

Long, Stephen, "Social Pressure and Contributions to Health Charities," *Public Choice*, Winter 1976, *31*, 55–66.

McGuire, Martin, "Group Size, Group Homogeneity and the Aggregate Provision of a Pure Public Good under Cournot Behavior," *Public Choice*, Summer 1974, *18*, 107–26.

Mackay, Robert J. and Whitney, Gerald A., "The Comparative Statics of Quantity Constraints and Conditional Demands: Theory and Applications," *Econometrica*, November 1980, *48*, 1727–43.

Margolis, Howard, "A New Model of Rational Choice," *Ethics*, January 1981, *91*, 265–79.

Moffitt, Robert, "The Effects of Grants-in-Aid on State and Local Expenditures: The Case of AFDC," *Journal of Public Economics*, April 1984, *23*, 278–306.

Paquè, Karl-Heinz, "Do Public Transfers 'Crowd Out' Private Charitable Giving? Some Econometric Evidence for the Federal Republic of Germany," Kiel Working Paper No. 152, August 1982.

Reece, William S., "Charitable Contributions: New Evidence on Household Behavior," *American Economic Review*, March 1979, *69*, 142–51.

______ **and Zieschang, Kimberly,** "Consistent Estimation of the Impact of Tax Deductibility on the Level of Charitable Contributions," *Econometrica*, March 1985, *53*, 271–93.

Roberts, Russell D., "A Positive Model of Private Charity and Public Transfers," *Journal of Political Economy*, February 1984, *92*, 136–48.

______, "A Taxonomy of Public Provision," *Public Choice*, No. 1, 1985, *47*, 267–303.

Romer, Thomas and Rosenthal, Howard, "Bureaucrats vs. Voters: On the Political Economy of Resource Allocation by Direct Democracy," *Quarterly Journal of Economics*, November 1979, *93*, 562–87.

Rose-Ackerman, Susan, "Do Government Grants to Charity Reduce Private Donations," in Michelle White, ed., *Nonprofit Firms in a Three Sector Economy*, Washington: Urban Institute, 1981.

______, "Charitable Giving and Excessive Fundraising," *Quarterly Journal of Economics*, May 1982, *97*, 195–212.

______, "Ideological Nonprofits and Government Grants: Some Thoughts on Nonprofits' Response to Federal Cutbacks," paper presented at the AEA annual meeting, New York, 1985.

Rudney, Gabriel, "A Quantitative Profile of the Nonprofit Sector," Working Paper No. 40, Program on Nonprofit Organizations, ISPS, Yale University, November 1981.

Schiff, Jerald, "Does Government Spending Crowd Out Charitable Contributions?," *National Tax Journal*, December 1985, *38*, 535–46.

Seaman, Bruce A., "Local Subsidization of Culture: A Public Choice Model Based on Household Utility Maximization," *Journal of Behavioral Economics*, Summer 1979, *8*, 93–131.

Steinberg, Richard, "Two Essays on the Nonprofit Sector," unpublished doctoral dissertation, University of Pennsylvania, 1983.

______, "Empirical Relations between Government Spending and Charitable Donations," *Journal of Voluntary Action Research*, Spring-Summer 1985, *14*, 54–64.

______, "Should Donors Care about Fundraising?," in Susan Rose-Ackerman, ed., *The Economics of Nonprofit Institutions: Studies in Structure and Policy*, New York: Oxford University Press, 1986.

Sugden, Robert, "Reciprocity: The Supply of Public Goods through Voluntary Contributions," *Economic Journal*, December 1984, *94*, 772–87.

Tobin, James and Houthakker, Henrik S., "The

Effects of Rationing on Demand Elasticities," *Review of Economic Studies*, No. 3, 1951, *18*, 140–53.

Tullock, Gordon, "Information without Profit," in D. M. Lamberton, ed., *Economics of Information and Knowledge*, Baltimore: Penguin Books, 1971.

Warr, Peter G., "Pareto Optimal Redistribution and Private Charity," *Journal of Public Economics*, October 1982, *19*, 131–38.

Weisbrod, Burton, "Toward a Theory of the Voluntary Non-Profit Sector in a Three Sector Economy," in Edmund Phelps, ed., *Altruism, Morality, and Economic Theory*, New York: Russell Sage, 1975.

Weiss, Jeffrey, "The Ambivalent Value of Voluntary Provision of Public Goods in a Political Economy," in Michelle White, ed., *Nonprofit Firms in a Three Sector Economy*, Washington: Urban Institute, 1981.

Wolpert, Julian, "Social Income and the Voluntary Sector," *Papers, Regional Science Association*, 1977, *39*, 217–29.

The Distribution of Public Services: An Exploration of Local Governmental Preferences

By JERE R. BEHRMAN AND STEVEN G. CRAIG*

A local governmental welfare function is specified to explore two of its central characteristics: the equity-productivity tradeoff and differential weights across neighborhoods. The model is estimated using service outputs (safety) in the welfare function, as opposed to publicly provided inputs (police), over neighborhoods. The equity-productivity tradeoff is found to be considerable, and not all neighborhoods are weighted equally. The estimation results raise several questions about accepted analysis of governmental behavior.

The *Serrano v. Priest* case concerning the allocation of educational expenditure in California brought the question of the nature of the distribution of local public services to the forefront of policy debate.[1] Despite considerable subsequent attention to related issues in the press and in political and judicial arenas, there has been little systematic economic analysis of important dimensions of this process. The fairly sparse related literature to date, recently surveyed by Edward Gramlich and Daniel Rubinfeld (1982), has focused on the empirical pro-poor vs. pro-rich bias of local public expenditures.

A problem with examining expenditures, however, is that what is of concern to residents is the actual level of service that is provided by local governments. The present paper therefore represents an important departure from most previous literature because we study distribution of local public service outcomes, rather than expenditure.[2] We hypothesize that local services are distributed "as if" there is a constrained maximization of a local governmental social welfare function,[3] defined over the distribution of local public services among the residents of its jurisdiction.[4] Differentiation between publicly provided inputs and final service outcomes reflects that there are two separate constraints on governmental welfare

*Professor of Economics, University of Pennsylvania, McNeil 160/CR, Philadelphia, PA 19104, and Assistant Professor of Economics, University of Houston, 4800 Calhoun Road, Houston, TX 77004. We thank Robert P. Inman and anonymous referees for useful comments, but the usual disclaimer applies. The crime survey is described in *Criminal Victimization Surveys*...(1976). This reference has a copy of the questionnaire, as well as descriptive statistics. Our data set was especially created for us by the Bureau of the Census.

[1] *Serrano* v. *Priest* (1971, L.A. 29820); subsequent opinion December 30, 1976. This is the landmark case in which the California Supreme Court ruled that the school finance system was unconstitutional due to the equal protection clause in the state constitution.

[2] This distinction itself is not original with us, though a number of studies seem to ignore it.

[3] The median voter model is the usual method for specifying local governmental preferences. However, this model has some well-known disadvantages, especially in the modeling the government of a large, heterogeneous city (see Robert Inman, 1979, for a discussion of these issues). Competing theories of governmental behavior are just beginning to be developed; they essentially involve group decision-making models (Kenneth Shepsle, 1979; Craig and Inman, 1985). These group behavior models may involve an agenda-setting politician, coalition building, or logrolling concensus building. We do not provide a structural model of governmental behavior, but we model the "as if" preferences of the government to allow for distributional concern, whatever the cause. We model local governmental preferences to depend on public services outcomes. Some observers suggest that expenditures (or, for constant prices across neighborhoods, inputs) may be arguments of a local governmental welfare function. Our approach is a start towards attempting to explain the observed distribution of publicly provided inputs given local governmental preferences over the service outcomes.

[4] This process operationalizes and significantly extends the local governmental choice framework first suggested, to our knowledge, by Carl Shoup (1964). We thank the referees for bringing Shoup's contribution to our attention.

maximization. The first is a resource constraint, which determines the amount of publicly provided inputs that are available. The second is a production constraint that determines how much public service outcome is produced by a combination of the publicly provided inputs and existing neighborhood characteristics.[5]

The distinction between publicly provided inputs and service outcomes is crucial because the allocation of inputs and the distribution of outcomes across neighborhoods can be very different. The difference can be illustrated in our model because we explicitly incorporate the production function constraint that converts inputs into outcomes. The model shows that the distribution of inputs (or expenditures) may respond in the opposite direction than the distribution of public service outcomes in response to changes in neighborhood characteristics. Further, we formulate the model to incorporate systematically different welfare weights for different neighborhoods; that may depend on characteristics such as income or racial composition.[6] We estimate the key parameters of our model using the allocation of police and safety from crime across neighborhoods in Baltimore. The empirical example shows how the distribution of final service outcomes depends on the key parameters of the "as if" welfare function.

The critical attributes of the local governmental welfare function are two. The first, inequality aversion, refers to the tradeoff between equity and productivity as reflected in the curvature of the welfare surface. The degree of curvature indicates the relative tradeoff between equity and concern over maximizing aggregate city-wide output (productivity). The second attribute of the welfare function, unequal concern, pertains to weights in the governmental welfare function for the service outcomes of different neighborhoods. Such weights may differ, for example, depending on neighborhood political support for the current local governmental incumbents, or on the possible movement of some residents from the jurisdiction to the detriment of the local tax base.[7] Unequal concern is reflected in the asymmetry of the welfare surface around a 45° ray from the origin.

Local government inequality aversion and unequal concern may underlie important differences between the publicly provided input and public service outcome distributions. A pro-poor distribution of publicly provided inputs across neighborhoods, for example, may result from a number of conceptually different phenomena:

1) The objective of the city government is to maximize aggregate service outcomes over the entire city, with no concern about distribution of those outcomes. Equivalently, there is no inequality aversion and there is no unequal concern. If publicly provided inputs would have a higher marginal product in poor neighborhoods were they distributed equally, resources are distributed in a pro-poor fashion for productivity reasons alone.

2) The objective is to equalize service outcomes for each neighborhood, in which case there is an extreme case of inequality aversion, but no unequal concern. If publicly provided inputs are more productive in rich neighborhoods, a pro-poor distribution of publicly provided inputs results because concern about equity overrides productivity considerations.

3) The objective of the city is to provide greater services for poor neighborhoods so

[5] Outcomes depend on both publicly provided inputs and private resident characteristics (see D. F. Bradford, R. A. Malt and W. E. Oates, 1969). For example, crime may be less in a "safe" neighborhood than in a "dangerous" neighborhood even if both have the same level of police activity.

[6] Distributional concerns in the policy arena are much broader than simply concern over income. Inman and Rubinfeld (1979), for example, show that racial concerns are potentially important in applying the *Serrano* decision to jurisdictions in which a distribution biased towards high-income groups would not be the basis for legal complaint.

[7] The group conflict models which underlie unequal concern imply many possible reasons why distribution may matter to local governments. Our specification is not a structural model of the causes of distributional concern, but it allows an empirical test of whether the group conflict models merit further investigation.

there is unequal concern favoring the poor. Even if publicly provided inputs are more productive in rich neighborhoods, a pro-poor distribution of such inputs may result.

These alternative scenarios illustrate that the distribution of *inputs* provided by a city government does not necessarily provide insight into the government's distributional interests regarding service *outcomes*. They also highlight the tradeoff faced by city government between equity and productivity and the possibility of unequal concern about different neighborhoods.

Our general model of the welfare-maximizing local government is presented in Section I. Section II discusses explicit functional forms and their implications. Section III presents the unique data set that we use, which involves the level of safety from crime in each neighborhood of a single jurisdiction. Section IV presents the empirical results. We find in the case of the distribution of police and of safety from crime among neighborhoods in Baltimore that the local government does sacrifice some productivity in order to achieve a more equitable distribution of service outcomes, and that unequal concern is pro-poor and pro-young, but racially neutral. A brief summary and conclusion is presented in the final section.

I. The Model

We assume that the local government acts as if it maximizes a welfare function defined over service outcomes in each neighborhood of the jurisdiction. In this one-period model, the resource constraint is assumed to be fixed. Further, the political structure is assumed constant, so the form of the welfare function also is exogenous. The model is developed specifically to account for the empirical example, the distribution of police and of safety from crime across neighborhoods of a single city. Nonetheless, the model is general to any public service outcome, and can be used to explore the allocation of any publicly provided input.[8]

The welfare function of concern to the local government is

$$(1)\qquad W=W(\underline{S},\underline{N}),$$

where $\underline{S}$ is a vector of outcomes such as safety from crime per capita in each of m neighborhoods and $\underline{N}$ is a vector of populations in each neighborhood.

The first derivatives with respect to both $\underline{S}$ and $\underline{N}$ are assumed to be positive. This welfare function is maximized subject to two constraints. First, there is a constraint on total governmental resources (R) which can be used to purchase public inputs, such as police:

$$(2)\qquad R\geq\sum_{j=1}^{m}TP_jN_j,$$

where R is the total available governmental resources (assumed to be fixed by the political process for the period of interest), $\underline{P}$ is a vector of per capita publicly provided inputs, where the jth element is the amount of the factor allocated to the jth neighborhood, and T is the price of P_j.

The second constraint specifies that production of the output, S, is dependent on the level of publicly provided inputs, P_j, and on a vector of neighborhood characteristics, $\underline{X}_j$.[9] The neighborhood characteristics in $\underline{X}_j$ are given for the period of the governmental allocation problem. Any neighborhood characteristics that adjust to the allocation of governmental resources within the time period of concern (for example, private security guards may be adjusted in response to police allocations) are not included in $\underline{X}_j$. Instead the private reaction functions for private inputs dependent on $\underline{X}_j$ and P_j are used to eliminate these inputs, so that S_j

[8] The model also is easily generalizable to a multitude of service outcomes and a multitude of publicly provided inputs. Here the one-outcome, one-input case is presented since that is the case explored in our empirical estimates. See Behrman (1986) for a multiple-input, multiple-output generalization for the intrahousehold allocation of nutrients.

[9] The neighborhood characteristics ($\underline{X}_j$) could include population or population density to capture contention effects or scale economies in the production of services. See Craig (1987b).

depends only on P_j and $\underline{X}_j$ in the following relation:[10]

$$S_j = f(P_j, \underline{X}_j), \tag{3}$$

where $\underline{X}_j$ is a vector of characteristics of the jth neighborhood that affect the outcome of interest but are not adjusted during the period.

We can obtain the first-order conditions under the assumptions that the welfare function in (1) and the production relation in (3) have the standard desirable properties for an interior maximum to occur. The intuition behind the model can be illustrated geometrically by considering the ratio of the first-order conditions for neighborhoods 1 and 2:

$$\frac{\partial W/\partial S_1}{\partial W/\partial S_2} = \frac{N_1\, \partial S_2/\partial P_2}{N_2\, \partial S_1/\partial P_1}. \tag{4}$$

The left side of (4) is the slope of the welfare function, for which W in Figure 1 indicates an iso-welfare curve.[11] The right side of (4) is the slope of the production possibility frontier, which is the convex solid line identified by S_1 and S_2. The production possibility frontier illustrates that the "price" of allocating more inputs to one neighborhood is the lost output in other neighborhoods. Welfare maximization leads to a tangency at point 1, at which point the marginal rate of substitution in the welfare function equals the marginal rate of transformation along the production possibility frontier.

In general, the slopes of the production possibility frontier for planes between different pairs of neighborhoods differ due to different values of $\underline{X}_j$ across neighborhoods (through equation (3)). Because the different pairwise planes of the production possibility frontier are tangent to the same welfare function, they trace out the curvature and location of that welfare function. For example, the dashed line in Figure 1 indicates the production possibility frontier in the plane for neighborhoods 3 and 4 with a tangency at point 2. The curvature of the welfare function can be identified by considering a series of points like 1 and 2. Note that estimation of relation (4) gives estimates of characteristics of the welfare function and not necessarily those of the production relation (3). Our analysis does account for the fact that both P_j and S_j may enter into relation (4), for which reason control for simultaneity is required.

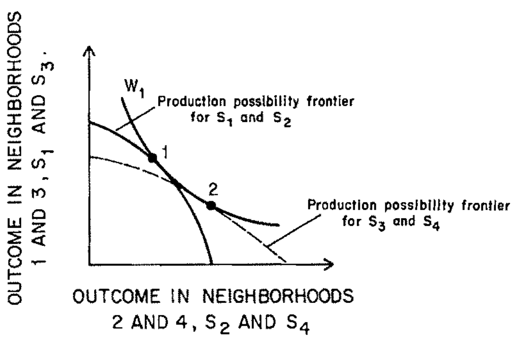

OUTCOME IN NEIGHBORHOODS 1 AND 3, S_1 AND S_3
W_1
Production possibility frontier for S_1 and S_2
1
2
Production possibility frontier for S_3 and S_4
OUTCOME IN NEIGHBORHOODS 2 AND 4, S_2 AND S_4

FIGURE 1. PRODUCTION POSSIBILITY FRONTIERS AND WELFARE-MAXIMIZING ALLOCATIONS

As we noted, the two critical attributes of the local governmental welfare function pertain to its curvature (inequality aversion) and to its asymmetry around a 45° ray from the origin (unequal concern). Figure 1 assumes some inequality aversion (with a curvature between the extreme linear case of focus only on productivity and the L-shaped case of focus only on equity) and equal concern. Equal concern about neighborhoods does *not* generally imply equal service outcomes across neighborhoods because the production set is not symmetrical if the distribution of $\underline{X}_j$ is not symmetrical. Figure 2 indicates a case of unequal concern in which neighborhood 2 is favored over neighborhood 1 in the sense that the weights in the welfare function are greater for neighborhood 2 than for neighborhood 1. Thus we distinguish between equal and unequal concern related to the *symmetry* of the iso-welfare curves around the 45° line and in-

[10] We are assuming a short-run allocation problem in which people do not move among neighborhoods because of the distribution of service outcomes. However, such movements could be made endogenous and dependent on S for a longer time horizon.

[11] This iso-welfare curve is drawn symmetrically around the 45° ray from the origin; it is not necessary to do so (see below).

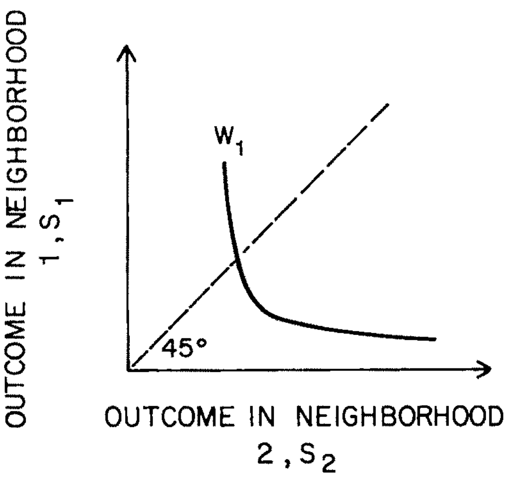

OUTCOME IN NEIGHBORHOOD 1, S_1
W1
45°
OUTCOME IN NEIGHBORHOOD 2, S_2

FIGURE 2. ILLUSTRATION OF UNEQUAL CONCERN ABOUT DISTRIBUTION OF THE OUTCOME

equality aversion (or the equity-productivity tradeoff) related to the *shape* of the iso-welfare curves. Both attributes of the welfare function are essential for determining the distribution of outcomes. For example, unequal concern even with pure inequality aversion (i.e., L-shaped iso-welfare curves) results in an unequal distribution of service outcomes.

II. Explicit Functional Forms and Relative vs. Absolute Inequality Aversion

This section presents explicit functional forms for the welfare function in (1), and for the production function constraint in (3). These functional forms allow estimation of the first-order conditions. From these results conclusions can be drawn as to what extent the local government trades off equity of public service outcomes for productivity (inequality aversion) and whether welfare weights differ across neighborhoods (unequal concern).

A. *Production Function Specification*

The production constraint on local governmental welfare maximization is assumed to have a partial log-linear form:

$$\ln S_j = \varepsilon \ln P_j + h(\underline{X}_j), \tag{5}$$

where ε is a production function elasticity, and h is any functional form with positive marginal productivities for $\underline{X}_j$.

The first-order condition for P_j is

$$\partial S_j / \partial P_j = \varepsilon S_j / P_j. \tag{6}$$

Since S_j depends on $\underline{X}_j$, this partial derivative depends on the neighborhood characteristics $\underline{X}_j$ and therefore changes as $\underline{X}_j$ changes, as is required for the identification of the welfare surface, even though $\underline{X}_j$ does not appear explicitly in (6).

B. *Welfare Function Specification*

We have explored two alternative specifications of the welfare function (1). One is the relatively well-known *CES* form; the other is the Kohm-Pollak (KP) specification (see Charles Blackorby and David Donaldson, 1980, and Behrman and Raaj Kumar Sah, 1984).[12] The difference in the specifications is in their treatment of the inequality aversion parameter. Empirically, we find more support for the KP specification, which is presented here. The *CES* specification is presented in the Appendix since this specification probably is more familiar and may be more appropriate than the KP form in other applications. The KP welfare function is

$$W^{KP} = \frac{1}{q} \ln \left[\sum_j \alpha_j \frac{N_j}{N} e^{qS_j} \right] \text{ for } \approx \tag{7}$$

where N is $\Sigma_j \alpha_j N_j$. The parameter q summarizes the equity-productivity tradeoff. The lower is q, the greater is governmental inequality aversion concerning the distribution of public services among neighborhoods. As $q \rightarrow 0$, the KP welfare function approaches the pure productivity sum over individual outcomes with no concern about inequality. For $q_i \rightarrow -\infty$, the KP welfare function ap-

[12] We also have considered a generalized *CES* form with displacement from the origin by a set of parameters b_j. These b_j a priori could relate to unequal concern. However, we find no empirical support for such displacements, so we do not consider them further in this paper.

proaches the L-shaped pure equity form. For values of q between these two extremes, there is an equity-productivity tradeoff.

The parameters α_j relate to equal vs. unequal concern. If there is equal concern, $\alpha_j = \alpha$ for all m neighborhoods. If there is unequal concern, α_j depends on neighborhood characteristics, such as racial composition or income level. The α_j parameters can be interpreted to control for the political influence of the jurisdiction, and therefore allow the parameter q to represent the direct equity vs. productivity tradeoff.

The representation of inequality aversion in the KP welfare function is absolute. This can be seen from calculating the shape of an isowelfare curve between neighborhoods 1 and 2:

$$(8) \qquad \left.\frac{dS_1}{dS_2}\right|_{W^{\mathrm{KP}}} = \frac{N_2\alpha_2}{N_1\alpha_1} e^{q(S_2 - S_1)}.$$

This relation says that along an iso-welfare curve it is the *absolute* difference in expected outcomes across neighborhoods (not their relative values as in the *CES* case in the Appendix) that is relevant.

The logarithm of the first-order condition for the constrained maximization of the KP welfare function (7) is

$$(9) \qquad \ln P_j = A^{\mathrm{KP}} + qS_j + \ln S_j + \ln \alpha_j$$

where

$$A^{\mathrm{KP}} = \ln\left(P\lambda \left[\sum_j \alpha_j N_j e^{qS_j} \right] \text{for} \approx / \varepsilon \right),$$

is a constant within a period, and λ is the Lagrangian multiplier for the budget constraint.

For empirical work, a stochastic term can be added to represent the fact that observed *ex post* outcomes differ from *ex ante* expected outcomes. Estimates then can be obtained of absolute inequality aversion (q) with data on outcomes and government-allocated inputs across neighborhoods. Without further assumptions or a priori information it is not possible to identify the absolute magnitudes of the components of A^{KP}. However, it is possible to identify whether the α_j reflect equal concern by substituting for α_j a relation dependent on neighborhood characteristics into (9). If neighborhood characteristics are found to influence the α_j, the hypothesis that all residents are weighted equally in the allocation process can be rejected.[13] Simultaneous estimation is required for (9) because it involved S_j and P_j, as does the production relation in (5).

III. Data: Safety from Crime and Police Allocation in Baltimore

Our empirical illustration of the model considers the allocation of police to produce safety from crime in the city of Baltimore in fiscal year 1972. The unit of observation is an individual neighborhood. The allocation of police in Baltimore is an interesting example for two reasons. First, there had been considerable local (Democratic) political stability in Baltimore at that time, with the same police chief (D. D. Pomerleau) since 1968, so the factors underlying the allocation processes probably are not masked by too much noise from adjustment processes. Second, a unique data base exists for Baltimore in 1972 which permits the estimation of our model, including the exploration of whether it is survey-reported crime or officially reported crime that matters in the allocation process.

The data base contains information on 79 (out of 240) representative residential neighborhood police beats.[14] It combines data from: 1) the *Criminal Victimization Surveys* on survey-reported crime per capita; 2) the

[13] The estimates of q are not particularly dependent on the choice of variables included in the α_j for our sample. The identification of the impact of neighborhood characteristics on α_j, however, is more problematic. See fn. 20 below.

[14] This includes all the police beats on which data were collected in the crime survey. The beats fairly represent neighborhoods in Baltimore (see Craig, 1987a, for more detail). The police beats included are not all contiguous. For this reason we have avoided the complications of including "spillover" effects on other neighborhoods in the models of Sections I and II; such an extension would be straightforward, but tedious.

Table 1—Data Definitions and Descriptive Statistics[a]

Variables	Means	Standard Deviations	Ranges
Outcomes[b]			
Per capita safety from officially reported crime	1.28	.28	.04–1.55
Per capita safety from survey reported crime	.54	.14	.09–.75
Officially reported crime per capita	.32	.28	.05–1.56
Survey reported crime per capita	.26	.14	.05–.71
Inputs			
Police patrols per capita ($\times 10^{-3}$)	1.010	.880	.095–5.195
Neighborhood Characteristics Tested for Relation to Unequal Concern			
Mean household income	\$6928	\$2941	\$2795–18179
Percent residents white	36.6	39.0	0–100
Percent residents over 65	13.1	8.4	0–35.2
Percent resident-owned housing	29.8	22.5	0–82.0
Other Instruments			
Percent households with income < \$5000	44.6	23.2	0–91.1
Percent households with income > \$15000	7.9	10.5	0–56.1
Percent single homes	60.8	26.7	0–100
Percent with ≥ 10 units per building	13.6	24.4	0–100
Percent married	39.1	14.3	0–69.4
Percent male	42.8	7.6	16.7–62.3
Percent between 16 and 24-years-old	21.3	7.7	0–44.4
Percent unemployed	2.4	3.6	0–21.6
Percent completed high school	85.9	16.6	19.3–100
Percent insured for loss	18.0	19.0	0–100
Percent criminals observed who are white	18.6	31.4	0–100
Percent crime victims employed	58.5	22.9	0–100
Average age of household head	38.4	6.1	25.6–53.2
Average dollar loss per crime	\$246.9	\$365.8	\$0–3000

[a] There are 79 neighborhoods in the sample. Data are for Baltimore in fiscal year 1972.

[b] Per capita safety from crime is defined as is indicated in relation (10): per capita safety from officially reported crime is 1.6 minus officially reported crime and per capita safety from survey reported crime is 0.8 minus survey-reported crime.

Baltimore Police Department on officially reported crime per capita and the number of police per capita; and 3) census data on neighborhood characteristics.[15] Table 1 gives summary statistics of the variables that we use.

We use two outcomes: safety from survey-reported crime per capita and safety from officially reported crime per capita. One interesting fact in the study of crime is that about one-half of all crime is unreported to the police. Because a survey measure of neighborhood crime is available, these data permit examination of whether the police authorities use officially reported or survey-reported crime in making police allocation decisions. Unfortunately, the survey measure of crime has some deficiencies; in particular, it does not measure crime to nonresidents in an area, while it includes crime to residents suffered in other areas. We restrict our analysis to residential crime to minimize these problems (see Craig, 1987a, for more details).

For both outcomes, we define safety from crime[16] to be

$$S_j = \bar{C} - C_j, \tag{10}$$

[15] The data base was prepared by Robert Highsmith while he was at Towson State University. We are grateful to him for kindly making it available for this study.

[16] Some outcomes may lessen "bads" such as crime (but perhaps in other contexts, disease) rather than increase "goods." Of course, goods can be considered to be the absence of bads. Because our functional forms use logarithms, for empirical purposes we use safety from crime as defined in relation (10) as the outcome of interest.

where $\bar{C} > \max(C_j)$ and C_j is the per capita crime rate (reported or actual) in the jth neighborhood. The empirical results are not qualitatively sensitive to a range of values for $\bar{C}$. We use a value slightly above $\max(C_j)$ in the estimates that we present below.[17]

The equation of interest for estimation is (9). The problem is a simultaneous one, as not only does the allocation of police depend on safety levels, but safety levels are produced by police. Therefore, we use an instrumental variable estimator, with a list of instruments that accounts for the supply of crime, the private demand for safety, the propensity to report crime, and exogenous determinants of the probability of arrest. The instruments are derived from a simultaneous equation model of urban crime; Craig (1987a) presents the full specification. These instruments include the 4 neighborhood characteristics indicated above, plus 14 additional characteristics (see Table 1). The basic thrust of our results is not sensitive to the exact instruments used.

IV. Estimates of Inequality Aversion and Unequal Concern in Baltimore Allocation of Police Patrols

The first set of empirical results that is presented utilizes a two-output version of the model to examine whether officially reported or survey-reported crime is the more relevant decision variable in the local government welfare function. Results are then presented for both the KP and *CES* specifications of the welfare function, although we concentrate on the empirically preferred KP specification.

A. *Safety from Officially Reported vs. Survey-Reported Crime*

As noted in Section III, our sample includes data on crime as reported in police records, and survey-reported crime as indicated in the National Crime Victimization Survey. We estimate a nonlinear relation parallel to the antilog of (9), but with an extension to the two-output case with safety from officially reported crime and safety from survey-reported crime as the two weighted outcomes, in order to examine which measure of public service outcome is more relevant empirically.[18]

TABLE 2—ESTIMATION OF KP WELFARE FUNCTION PARAMETERS WITH TWO OUTCOMES: OFFICIALLY REPORTED AND SURVEY-REPORTED CRIME[a]

Right-Side Variables	Parameters in Two-Outcome Extension of First-Order Condition in Relation (9)
Multiplicative Coefficient of[b]	
Safety from Officially Reported Crime	45.1[d] (14.9)
Safety from Survey-Reported Crime	−5.6 (5.4)
Inequality Aversion Parameter[c]	
Safety from Officially Reported Crime	−2.9[d] (.51)
Safety from Survey-Reported Crime	−3.0 (1.7)
R^2	.76
SEE	27.5

[a] Beneath the point estimates are the standard errors of estimates. The R^2 are pseudo R^2 calculated as one minus sum of squares of the error over the total sum of squares of the dependent variable.

[b] The first two rows give the weights on the safety from reported and survey crime terms, respectively.

[c] The third and fourth rows give the estimates for the inequality aversion parameter (i.e., q in (9)).

[d] Point estimates that are significantly nonzero at the 5 percent level.

Since the results are robust with regard to variations in specification, we report in Table 2 the estimates for the simplest case—the basic KP model with equal concern. We find that the weights on safety from officially reported crimes are positive and significant while those for safety from survey-reported crime are negative and insignificant.

There are at least two possible explanations for finding that officially reported crime

[17] To be explicit, we use $\bar{C} = 1.6$ for per capita safety from officially reported crime and $\bar{C} = 0.8$ for per capita safety from survey-reported crime.

[18] In such an estimation the weights cannot be identified separately from the product of the equal concern parameters (α_j), the constants (A^{KP}), and the production function parameter (ε). But the significance of the product that includes the weights can be determined from the estimates.

TABLE 3—KP AND *CES* WELFARE FUNCTION PARAMETER ESTIMATES WITH AND WITHOUT EQUAL CONCERN[a]

	Absolute Inequality Aversion KP—Equation (9)		Relative Inequality Aversion *CES*—Equation (A2)	
Right-Side Variables	Equal Concern (1)	Unequal Concern (2)	Equal Concern (3)	Unequal Concern (4)
Estimates of Inequality Aversion:				
ln Safety from Crime	1.0[b]	1.0[b]	−1.6[c]	−.90[c]
per Capita (*c* for *CES*)			(.30)	(.30)
Safety per Capita (*q* for KP)	−4.0[c]	−3.4[c]		
	(.30)	(.39)		
Determinants of Unequal Concern:				
α (or *a*)				
Mean household income		−.55[c]		−.66[c]
		(.17)		(.20)
Percent residents white		−.02		−.02
		(.02)		(.02)
Percent residents over 65		−.06[c]		−.08[c]
		(.03)		(.04)
Percent resident-owned housing		−.01		−.06[c]
		(.03)		(.03)
Constant	4.5[c]	8.8[c]	−.04	6.0[c]
	(.38)	(1.3)	(.10)	(1.7)
R^2	.53	.67	.27	.55
SEE	24.7	17.1	48.3	25.8

[a] See fn. a, Table 2.

[b] In these estimates the coefficient of ln safety is constrained to be one as required in relation (9).

[c] See fn. d, Table 2.

empirically dominates as a determinant of the allocation process. First, the officially reported information may be the best (or only) information available to the allocator of police across neighborhoods. Even if the survey crime measure is better, the police may not have accurate information on how officially reported and actual crime diverge systematically across neighborhoods. Second, even if the allocator of police knows the pattern of actual in addition to officially reported crimes and survey-reported crime is a better measure of actual crime than is officially reported crime, attention may be focused on officially reported crimes because of the perception that they are more important in the political process.[19] If either of these explanations hold, there may be social gains in terms of control of total crime from improving the data on officially reported crime to reflect better the patterns of actual crime.

In any case, because of this evidence on the relative importance of safety from officially reported crime in the police allocation process, we focus exclusively on officially reported crime in what follows.

B. *Welfare Parameters Underlying Police Allocation*

Table 3 presents estimates of specifications with equal and with unequal concern for the first-order conditions derived from the KP (9) and *CES* (A2) welfare functions. These results are robust in three crucial respects. First, and most important, they indicate substantial governmental inequality aversion, but still with some equity-productivity tradeoff. Second, they indicate that

[19] See Craig (1987a) for additional discussion of this issue, an estimate of the causes of the propensity to report crime, and a discussion of the problems with the survey crime measure.

unequal concern prevails in the allocation decisions, so that all neighborhoods are not weighted equally in the "as if" social welfare function. Third, they suggest that police patrol allocations are more consistent with the KP absolute inequality aversion than with the *CES* relative inequality aversion local governmental welfare function (see the Appendix for the *CES* specification).

The statistically significant inequality aversion is in addition to any unequal distribution of inputs resulting from different social welfare weights for different neighborhoods. The KP results allowing for unequal concern yield an estimate of q of -3.4. This estimate is significantly less than zero, indicating aggregate outcome is not maximized. At the same time, however, there also is significant concern about productivity, as the inequality aversion parameter is significantly greater than the extreme of concern solely with equity. Results for the *CES* case are similar; the estimated value of c is significantly negative with both the equal and the unequal concern specifications.

In addition to the inequality aversion, *F*-tests reject the hypothesis of equal concern across neighborhoods.[20] As discussed in Section II, this means that public service output is weighted more heavily in the local governmental welfare function for some neighborhoods than for others. The estimates with unequal concern are robust in suggesting preferences for greater safety from crime for lower income and younger (in the sense of a smaller percentage of residents over 65) neighborhoods, but there is no significant impact of racial composition. The estimates are mixed for the percentage of resident-owned housing, with a significant negative effect in the *CES* case, but insignificant impact in the preferred KP case. Thus the preference weights seem to be pro-poor and pro-young, but neutral regarding race and probably resident-owned housing.

Some important implications of our estimates of inequality aversion and of unequal concern can be seen by solving relations (5) and (9) for the reduced form for the ratio of safety from crime:[21]

$$(11)\quad \frac{e^{h(\underline{X}_1)}}{e^{h(\underline{X}_2)}}\left(\frac{\alpha_1}{\alpha_2}\right)^{\varepsilon}=\left(\frac{S_1}{S_2}\right)^{1-\varepsilon}\frac{e^{-qS_1\varepsilon}}{e^{-qS_2\varepsilon}}.$$

Since $0<\varepsilon<1$ (because it is the elasticity of safety from crime with respect to police patrols) and q is estimated to be less than zero (see Table 3), all of the powers on the right side of relation (11) are positive; therefore, if neighborhood 1 has more of the characteristics that produce safety from crime and there is equal concern ($\alpha_1=\alpha_2$), neighborhood 1 has greater safety ($S_1>S_2$).

Relation (11) also is useful for examining the optimal allocation of safety from crime in neighborhood 1 vs. 2. The more unequal concern favors neighborhood 1 (i.e., the greater is α_1/α_2), the greater is safety from crime in neighborhood 1 relative to 2. Since the mechanism for increasing safety from crime in neighborhood 1 vs. 2 is through allocation of police patrols (given neighborhood characteristics), relation (11) also implies that the more unequal concern favors neighborhood 1, the greater the relative al-

[20] Estimation of a model that allows unequal concern may have a greater problem with identification than does estimation of a model assuming equal concern because of the introduction of neighborhood characteristics. We have noted in Sections I and II that our estimates are of first-order conditions that permit identification of certain characteristics of the welfare function and *not* of production relations, which also involve the neighborhood characteristics in $\underline{X}_j$. If the neighborhood characteristics in $\underline{X}_j$ were exactly the same as the neighborhood characteristics on which unequal concern depends, however, our first-order conditions would be identified from production relations only by the functional forms (and, as a referee has noted, some alternative production relation to that in (5) may make identification by functional form impossible). On a priori grounds we do not think that the set of neighborhood characteristics that we explore in our estimates are identical to the elements of $\underline{X}_j$ in the production relation (also see Craig, 1987a). Moreover, our estimates of the inequality aversion parameters do not change significantly if we allow unequal concern, as might be expected were we switching from estimating the first-order condition to the production function. Therefore we interpret our estimates to relate to unequal concern and not to the production process.

[21] Closed-form expressions for the ratio of police patrols and safety from crime cannot be derived for the KP case. See the Appendix for such expressions in the *CES* case.

location of police patrols to that neighborhood.

These estimates imply an allocation of police resources that partially compensates for the distribution of neighborhood characteristics that prevent crime. While there is ambiguity in the KP case about which neighborhoods have greater police patrols, relation (11) does show that the relative allocation of police is greater to neighborhoods with more crime-causing characteristics. But compensation does not offset completely the impact of neighborhood characteristics, so neighborhoods with more crime prevention characteristics are more safe than are those with relatively greater police patrols. In addition, those neighborhoods that are favored by unequal concern are allocated further police and have greater safety from crime than would be the case with equal concern.

V. Summary and Conclusions

We have presented a model which explicitly examines the distributional preferences of a local government over service outcomes. An important feature of the model is that it distinguishes inequality aversion from unequal concern in the social welfare function of a local government. The model is estimated for a particular case, the allocation of safety from crime and of police across neighborhoods in Baltimore in 1972. All the major results hold with both alternative specifications of the welfare function. These empirical results indicate a significant degree of inequality aversion, so that some aggregate production is sacrificed in order to obtain the equity goals of the local government. Further, there appears to be unequal concern so that the safety from crime outcomes are weighted differently for residents in different neighborhoods. Subject to qualifications about the conditionality of our estimates on our assumptions and data, our empirical results have potentially far-reaching implications for models of local government behavior.

First, we have shown that at least one major local government appears to have substantial concern about equity in the distribution of a local public service. The inequality-aversion parameters that we estimate are significantly negative, even after accounting for the fact that residents may be weighted unequally in the social welfare function. This suggests that the local government compensates for the distribution of characteristics across neighborhoods regarding safety, rather than reinforcing the impact of such characteristics, in its allocation of police across neighborhoods. While a structural model that would explain strong inequality aversion has not yet been developed, our contribution shows that pursuing such a research goal may yield interesting new insights into governmental behavior.

Second, we also show that in this case not all residents are weighted equally in the local welfare function. These results also merit further examination. Unequal concern may exist, for example, because residents who receive less safety from crime may receive more of other publicly provided services, such as education. Conversely, the results may reflect that certain residents are more "in favor" with the authorities, and receive more of all locally provided services. While our single-input, single-outcome estimates are not able to distinguish between these two alternatives, or from a host of other explanations, the point is that a distinction among residents apparently is being made by the government. Research with a multi-output extension of our model, if data become available, could clarify further the situation. In any case, our results imply that models which a priori assume that all residents are treated equally in the distribution of governmental services are ignoring the potentially important fact that unequal welfare weights may prevail.

Third, our empirical support in this case for both aspects of unequal service allocation by neighborhood implies potentially serious misspecification in studies of aggregate local public service demand. Models of preference aggregation may need to take heed of the fact that people pay taxes based on the city-wide amount of purchased inputs, but base their demand and voting behavior on the perceived level of neighborhood service output. The standard assumption of median voter models is that residents participate equally in service outputs and tax shares. These models calculate resident tax shares based upon their share in the cost of the aggregate level of purchased inputs. However, the framework and results presented here show that residents may share unequally in the benefit of those inputs to the extent that service outcomes differ from the allocation of inputs. Thus, residents may perceive different service levels even when they have equal tax shares, causing differences in the voting behavior of people who prima facie may appear to face the same constraints.

Appendix

The CES Welfare Function Specification

The *CES* specification of the welfare function is

$$\text{(A1)} \qquad W^{CES} = \left(\sum_j N_j a_j S_j^c\right)^{1/c}.$$

The parameter c refers to inequality aversion. Like q for the KP specification, as c is more negative, inequality aversion is greater. At the extreme with only concern about equity, c is $-\infty$; for the intermediate Cobb-Douglas case, c is zero; and for the extreme with only concern about productivity, c is one so that W^{CES} is the weighted sum of the S_j. The a_j parameters represent unequal concern, in the same manner as the α_j parameters in the KP version. Inequality aversion in the *CES* case is relative (i.e., along an isowelfare curve it is the *relative* outcomes that matters), rather than absolute as in the KP case. Nonetheless, the estimation results are similar for the two cases (see Table 3).

Maximization of (A1) subject to the resource and production constraints yields an estimating equation from the first-order conditions that is similar to (9) for the KP case:

$$\text{(A2)} \qquad \ln P_j = A^{CES} + c \ln S_j + \ln a_j,$$

where

$$A^{CES} = \ln\left(P\lambda \Big/ \left(\varepsilon\left(\sum_j N_j a_j S_j^c\right)^{(1-c)/c}\right)\right)$$

is a constant within a period.

One advantage of the *CES* case is that closed-form expressions can be derived for the ratios of P_1/P_2 and S_1/S_2 for neighborhoods 1 and 2, analogous to equation (11) for the KP specification:

$$\text{(A3)} \qquad \frac{P_1}{P_2} = \left(\frac{a_1}{a_2}\right)^{1/(1-\varepsilon c)} \left(\frac{e^{h(\underline{X}_1)}}{e^{h(\underline{X}_2)}}\right)^{c/(1-\varepsilon c)};$$

$$\text{(A4)} \qquad \frac{S_1}{S_2} = \left(\frac{a_1}{a_2}\right)^{\varepsilon/(1-\varepsilon c)} \left(\frac{e^{h(\underline{X}_1)}}{e^{h(\underline{X}_2)}}\right)^{1/(1-\varepsilon c)}.$$

Again, the implications of these expressions are similar to those for the KP version, except there is no ambiguity; the allocation of police patrols is greater to the neighborhood with less crime prevention characteristics for our estimate of c in Table 3.

REFERENCES

Behrman, Jere R., "Intrahousehold Allocation of Nutrients in Rural India: Are Boys Favored? Do Parents Exhibit Inequality Aversion?," mimeo., University of Pennsylvania, 1986.

______ **and Sah, Raaj Kumar,** "What Role Does Equity Play in the International Distribution of Aid?," in Moises Syrquin et al., eds., *Economic Structure and Performance*, New York: Academic Press, 1984, 295–315.

Blackorby, Charles and Donaldson, David, "A Theoretical Treatment of Indices of Absolute Inequality," *International Economic Review*, February 1980, *21*, 107–136.

Bradford, D. F., Malt, R. A. and Oates, W. E., "The Rising Cost of Local Public Services," *National Tax Journal*, June 1969, *22*, 185–202.

Craig, Steven G., (1987a) "The Deterrent Impact of Police: An Examination of a Local

Public Good," *Journal of Urban Economics*, forthcoming 1987.

______, (1987b) "The Impact of Congestion on Local Public Good Production," *Journal of Public Economics*, forthcoming 1987.

______ **and Inman, Robert P.**, "Education, Welfare and the 'New' Federalism: State Budgeting in a Federalist Public Economy," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1985, 187–221.

Gramlich, Edward M. and Rubinfeld, Daniel L., "Micro Estimates of Public Spending Demand Functions and Tests of the Tiebout and Median-Voter Hypotheses," *Journal of Political Economy*, June 1982, *90*, 536–60.

Inman, Robert P., "The Fiscal Performance of Local Governments: An Interpretative Review," in Peter Mieszkowski and Mahlon Straszheim, eds., *Current Issues in Urban Economics*, Baltimore: Johns Hopkins University Press, 1979, 270–321.

______ **and Rubinfeld, Daniel L.**, "The Judicial Pursuit of Local Fiscal Equity," *Harvard Law Review*, June 1979, *92*, 1661–750.

Shepsle, Kenneth, "Institutional Arrangements and Equilibrium in Multidimensional Voting Models," *American Journal of Political Science*, January 1979, *23*, 27–59.

Shoup, Carl S., "Standards for Distributing a Free Governmental Service: Crime Protection," *Public Finance*, No. 4, 1964, *19*, 383–92.

U.S. Department of Justice, *Criminal Victimization Surveys in Eight American Cities*, No. SD-NCS-C-5, Washington: Law Enforcement Assistance Administration, November 1976.

Intertemporal Labor Supply and Long-Term Employment Contracts

By JOHN M. ABOWD AND DAVID CARD*

We compare a contracting model and a labor supply model. One test is whether earnings changes are more variable than hours changes, as predicted by the labor supply model, or less variable, as predicted by the contracting model. We apply this test to two longitudinal surveys and find that earnings are somewhat more variable than hours for men who never change employers. The estimates suggest that changes in earnings and hours not associated with measurement error occur at fixed wage rates.

Despite rapid progress over the last decade in modeling employment contracts and recent evidence on the importance of long-term jobs in the economy, microeconomic studies of labor supply continue to interpret individual hours and earnings data in terms of an auction model of the labor market.[1] Traditional labor supply models assume that earnings represent the product of desired hours and market wage rates. Contracting models, on the other hand, interpret earnings as optimal consumption for the payment period, including savings and insurance payments from firms to workers.[2] If savings and insurance are important components of earnings, then average hourly earnings provide, at best, noisy information on underlying productivity. Contract models, therefore, offer a simple explanation for the weak link between wage rates and hours that has confounded empirical studies of intertemporal labor supply.[3]

In this paper we compare the implications of a life cycle labor supply model and an intertemporal contracting model for changes in individual earnings and hours over time. Specifically, we compare a dynamic labor supply model in which individuals have access to complete capital markets to a symmetric information contracting model in which employees receive complete insurance from their employers. The critical distinction between these models is whether earnings represent optimal consumption or the product of wage rates and hours of work. We develop a simple test between the two models based on the relative variability of earnings and hours with respect to changes in productivity. If earnings represent the product of wages and hours, then changes in productivity generate bigger changes in earnings than hours. If earnings represent consumption, on the other hand, then changes in productivity generate smaller changes in earnings than hours, provided that leisure is a normal good.

This simple test is complicated by changes in earnings and hours that may occur with movements across jobs. Although the labor

*Department of Economics, Princeton University, Princeton, NJ 08544. We acknowledge the helpful comments of Gary Chamberlain, Henry Farber, Zvi Griliches, and John Kennan, on earlier drafts. We are also grateful to the Department of Economics at the Massachusetts Institute of Technology for their hospitality while this paper was completed. This paper was originally circulated under the title "Intertemporal Labor Supply in the Presence of Long Term Contracts." Our research was partially supported by the University of Chicago Graduate School of Business.

[1] The recent literature on labor supply is surveyed by Mark Killingsworth (1983) and John Pencavel (1987). The long duration of jobs is emphasized by Robert Hall (1982). Estimates of completed job durations for adult males are presented by Katharine Abraham and Henry Farber (1985).

[2] Oliver Hart (1983) and Sherwin Rosen (1985) provide useful surveys of the literature on contracting models.

[3] This point is emphasized by Rosen (1985). James Brown (1982) estimates aggregate employment equations that include savings and insurance components in earnings.

supply model makes no distinction between changes within and across jobs, the contracting model is employer-specific. We therefore propose the following test of the two alternative models: compare the relative contribution of productivity shocks to changes in earnings and changes in hours for workers who are observed on the same job over time. If, as the intertemporal contracting model suggests, these workers are fully insured, then the contribution of productivity shocks to changes in earnings should be smaller than the contribution of productivity shocks to changes in hours. If the labor supply model is correct, on the other hand, then the contribution of these shocks to changes in earnings should be greater than the contribution to changes in hours.

Our empirical analysis is conducted with data from the *Panel Study of Income Dynamics* (*PSID*) and the *National Longitudinal Survey of Men 45–59* (*NLS*). In order to identify workers who are potentially covered by long-term contracts, we distinguish between individuals who report the same employer during the sample period, and individuals who change employers at least once. We find that earnings and hours changes are substantially less variable for individuals who do not change employers. For both groups of workers, the contribution of productivity shocks to earnings is *greater* than the contribution to hours, although we cannot reject the hypothesis that earnings and hours move proportionately with changes in productivity. This finding casts doubt on the usefulness of either consumption-smoothing contract models or dynamic labor supply models. In fact, a simple interpretation of the data is that earnings and hours vary at fixed hourly wage rates.

Since the focus of this paper is on the contrast between labor supply and contracting models of earnings and hours variation, we concentrate on relatively simple specifications of the components of earnings and hours. In our other work (1986), we investigate more general factor-analytic models of the covariance structure of earnings and hours changes. Although these models provide a somewhat better description of earnings and hours changes, we find that for adult male workers, most systematic hours variation occurs at fixed wage rates. The conclusions in this paper, therefore, are not affected by extensions to the simple components of variance models of earnings and hours presented here.

Section I presents a simple theoretical analysis of intertemporal contracting and intertemporal labor supply models. For both models we derive the implications of changes in productivity for relative changes in earnings and hours. These implications provide the basis for our empirical test between the models.

In Section II, we show how to estimate the theoretical models using the variances, autocovariances, and cross-covariances of earnings and hours changes from individual longitudinal data. A two-factor variance components model provides a convenient framework for distinguishing changes in productivity from other sources of earnings and hours variation, including changes in tastes and measurement error.

Section III summarizes the data from both surveys and presents estimates of the structural parameter that distinguishes the contracting and labor supply models. The covariance structure of earnings and hours changes is remarkably similar in the two surveys. Our main empirical finding is that productivity variation affects earnings at least as much as hours. This is true for individuals who have the same employer in all years and for those who change employers. The data therefore provide some evidence against a contracting interpretation. They also suggest, however, that productivity-related changes in earnings and hours occur at fixed wage rates.

I. Earnings and Hours under Long-Term Contracting Models and Life Cycle Labor Supply Models

Here we present a simple dynamic model of earnings and hours determination under long-term employment contracts.[4] We also

[4] The modern analysis of implicit contracts begins with Walter Oi (1962) and Rosen (1968). The macroeco-

present a model of earnings and hours determination under a standard life cycle labor supply framework.[5] We make identical assumptions about preferences and individual productivity in the two models. For the contracting model, we assume that employers have access to complete capital and insurance markets. For the labor supply model, we assume that individuals have direct access to these markets. Our models, therefore, contrast a widely used version of the intertemporal labor supply model with a class of testable contracting models.

Individual productivity is modeled as a random variable drawn from a sequence of distributions that are common knowledge for both workers and firms. Productivity is the only source of uncertainty in the model. Apart from firm-specific training and recruiting costs, individuals are equally productive at all firms. Long-term attachments between workers and firms arise from two sources: first, the desire to avoid recurrent training costs, which occurs in either the contracting or labor supply model; and, second, the desire to smooth consumption vis-à-vis productivity, which is associated with long-term attachments in the contracting model.

Preferences for consumption and leisure within periods are modeled as a general function of consumption, hours of work, and age. Preferences are assumed to be additively separable over time and across states of productivity. The worker's intertemporal objective is to maximize the expected discounted value of lifetime utility. In the contracting model, the expectation is taken over the distribution of individual productivities. In the labor supply model, the expectation is taken over the distribution of market wages, which is assumed to be identical to the distribution of individual productivities.

Let θ_t represent the productivity of a given individual in period t.[6] Assume that θ_t is distributed on the interval (θ_l, θ_u) according to a known distribution function $F_t(\theta_t)$. Let $u(c_t(\theta_t), h_t(\theta_t), t)$ represent a concave von Neumann-Morgenstern utility function over consumption (c) and hours of work (h) in period t. Let the utility discount rate be ρ. The worker's objective is to maximize expected utility denoted by

$$(1) \quad \sum_{t=0}^{T}\left(\frac{1}{1+\rho}\right)^{t} \times \int_{\theta_l}^{\theta_u} u(c_t(\theta_t), h_t(\theta_t), t)\, dF_t(\theta_t),$$

where T represents a fixed planning horizon.

Consider the long-term contracting model first. Firms offer contracts consisting of contingent labor demand functions $h_t(\theta_t)$ and contingent earnings functions $g_t(\theta_t)$ for $t = 1,\ldots,T$. Since workers have no access to capital markets, $g_t(\theta_t) = c_t(\theta_t)$ for all t. If productivity is θ_t in period t, the firm's revenues are $\theta_t h_t(\theta_t)$ and its costs are $g_t(\theta_t)$. We assume that θ_t is observable and, therefore, contracts are fully enforceable. We also assume that firms are risk neutral and can borrow and lend at the constant real interest rate r.[7] Competition among firms for the services of a worker with the sequence of productivity distributions $\{F_t(\theta_t)\}$ implies that contracts offered to that worker have expected present value equal to the training

nomic implications of employment contracts were emphasized by Martin Baily (1974), Donald Gordon (1974), and Costas Azariadis (1975).

[5] Studies of life cycle labor supply and consumption originate in Franco Modigliani and Richard Brumberg's (1954) analysis of the divergence between planned consumption and earnings over the life cycle and in Milton Friedman's (1957) study of the consumption function. Robert Lucas and Leonard Rapping (1969) use a two-period model in their influential study of intertemporal substitution and labor supply. Multiperiod labor supply is considered by James Heckman (1974, 1976), Gilbert Ghez and Gary Becker (1975), and many subsequent authors, in particular, Thomas MaCurdy (1981).

[6] For notational simplicity, we suppress the dependence of θ_t on the individual. Randomness of θ_t is over *ex ante* identical individuals.

[7] Implicitly we are assuming that productivity risks are fully diversifiable. See Rosen (1985, pp. 1153–54) for a discussion of aggregate vs. idiosyncratic productivity risks and the implications of nondiversifiability.

costs, R, for that worker:

$$(2)\quad \sum_{t=0}^{T}\left(\frac{1}{1+r}\right)^{t} \times \int_{\theta_t}^{\theta_u}[\theta_t h_t(\theta_t) - g_t(\theta_t)]\, dF_t(\theta_t) = R.$$

Pointwise optimization of the Lagrangian expression for the maximization of (1), subject to (2), leads to the first-order conditions:

$$(3a)\quad \left(\frac{1+r}{1+\rho}\right)^{t} u_c(c_t(\theta_t), h_t(\theta_t), t) - \lambda = 0,$$

$$(3b)\quad \left(\frac{1+r}{1+\rho}\right)^{t} u_h(c_t(\theta_t), h_t(\theta_t), t) + \lambda\theta_t = 0,$$

where u_c and u_h represent the partial derivatives of $u(\cdot,\cdot,\cdot)$ with respect to c and h, and λ represents the multiplier associated with the constraint (2). Equations (3a) and (3b) have the familiar implications that the marginal utility of consumption follows a deterministic trend, while the marginal rate of substitution between consumption and leisure equals θ_t for each realization of productivity.

Differentiation of the first-order conditions (3a) and (3b) and some rearrangement yields

$$(4)\quad \frac{\partial \log h_t}{\partial \log \theta_t} - \frac{\partial \log h_t}{\partial \log \nu_t} = \frac{c_t}{\theta_t h_t}\frac{\partial \log c_t}{\partial \log \theta_t},$$

where $\nu_t \equiv \lambda((1+\rho)/(1+r))^t$. To understand the implications of equation (4), consider the log-linear approximation to the solution of equations (3a) and (3b):

$$(5a)\quad \log c_t = \phi \log \theta_t - \alpha \log \nu_t + a_t,$$

$$(5b)\quad \log h_t = \eta \log \theta_t + \delta \log \nu_t + b_t,$$

where a_t and b_t are time-varying terms in the log-linear approximation representing shifts in tastes for consumption and leisure. The parameter ϕ represents the substitution elasticity between consumption and leisure holding constant the marginal utility of wealth: the sign of ϕ depends on the sign of u_{ch}. If the permanent income hypothesis is correct, for example, then $\phi = 0$ and consumption is independent of productivity. The parameter η represents the elasticity of substitution of labor supply over time and across states of θ; therefore, $\eta \geq 0$. The parameter $-\alpha$ represents the elasticity of consumption demand with respect to the marginal utility of wealth; if consumption is a normal good, then $\alpha \geq 0$.[8] Finally, the parameter δ represents the elasticity of labor supply with respect to the marginal utility of wealth; if leisure is a normal good, then $\delta \geq 0$. Since $E[c_t] \cong E[\theta_t h_t]$ by constraint (2),[9] the restriction (4) implies (to a first-order approximation) $\eta - \delta \cong \phi$, or

$$(6)\quad \mu \equiv \phi/\eta \cong 1-(\delta/\eta).$$

The parameter μ represents the relative sensitivity of consumption and hours choices to changes in productivity. Even in the absence of direct information on productivity, μ is identifiable from information on the relative variability of earnings and hours. If $\mu \geq 1$, then $\delta \leq 0$; that is, if consumption is more variable than hours with respect to changes in productivity, then leisure is an inferior good. If $\delta > 0$ is treated as a maintained hypothesis, then the intertemporal contracting model presents one testable implication: namely, that changes in productivity influence earnings *less* than hours, on average.

Now consider the intertemporal labor supply model. We assume that workers have access to risk-neutral insurance and capital markets so that the life cycle budget con-

[8] A concave utility function implies that λ is a decreasing function of wealth, and therefore that the sign of the derivative of the demand for consumption goods with respect to λ is the same as the sign of the derivative of demand for consumption goods with respect to income.

[9] This holds for small training costs, R.

straint can be replaced by its expectation:[10]

$$(7)\quad \sum_{t=0}^{T}\left(\frac{1}{1+r}\right)^{t} \times \int_{\theta_l}^{\theta_u}\left[\theta_t h_t(\theta_t)-c_t(\theta_t)\right] dF_t(\theta_t)=0.$$

The first-order conditions for the maximization of (1) subject to the constraint (7) are identical to (3a) and (3b). Labor earnings, however, are now described by $g_t(\theta_t)=\theta_t h_t(\theta_t)$. The log-linear form of the solution for g_t and h_t becomes

$$(8a)\quad \log g_t=(1+\eta)\log\theta_t+\delta\log\nu_t+b_t;$$

$$(8b)\quad \log h_t=\eta\log\theta_t+\delta\log\nu_t+b_t,$$

where, as before, b_t represents a time-varying component of tastes for leisure. Under the labor supply interpretation, the variability of earnings relative to hours with respect to changes in productivity is given by

$$(9)\qquad \mu\equiv(1+\eta)/\eta.$$

Since earnings represent the product of wages and hours in the labor supply model, earnings must respond *more* than hours to changes in productivity.

Our analysis of the contracting model shows that the elasticity of earnings with respect to productivity is less than the elasticity of hours with respect to productivity if leisure is a normal good. Our analysis of the intertemporal labor supply model shows that the relation between these elasticities is reversed under identical assumptions.[11] In the next section we develop a statistical model for estimating the critical parameter μ, the ratio of the two elasticities.

II. Econometric Models for the Covariance Structure of Earnings and Hours Changes

Our empirical strategy is to fit the model of earnings and hours implied by the contracting model (equations (5a) and (5b)) and the labor supply model (equations (8a) and (8b)) to estimated covariance matrices of earnings and hours changes from longitudinal survey data. We develop a two-factor interpretation of earnings and hours changes that allows us to distinguish between productivity components of earnings and hours, on one hand, and components of variance associated with preference variation and survey measurement error, on the other. The two-factor structure permits direct estimation of μ, the relative variability of earnings and hours with respect to productivity changes, as well as an overall measure of the goodness of fit of this simple class of models to the covariance structure of earnings and hours changes.

For the contracting model, the first step is to express equations (5a) and (5b) in first-difference form, taking account of individual-specific components. Since earnings are identical to consumption in this model, we substitute $\log g_t$ for $\log c_t$. Let $\Delta\log g_{it}$ and $\Delta\log h_{it}$ represent the observed changes in the logarithms of real annual earnings and annual hours for individual i between periods $t-1$ and t, respectively. Append a survey measurement error u_{it}^* to the expression for $\log g_{it}$ and a survey measurement error v_{it}^* to the expression for $\log h_{it}$. Then, equations (5) imply

$$(10a)\quad \Delta\log g_{it}=\phi\Delta\log\theta_{it} -\alpha(\rho-r)+\Delta a_{it}+\Delta u_{it}^*,$$

$$(10b)\quad \Delta\log h_{it}=\eta\Delta\log\theta_{it} +\delta(\rho-r)+\Delta b_{it}+\Delta v_{it}^*.$$

These equations express observed changes in earnings and hours in terms of changes in productivity, changes in tastes for consumption and leisure, and changes in measurement error. Since employees can perfectly insure individual productivity variation, the

[10] Robert Topel and Finis Welch (1986) refer to this model as a contracting model with self-insurance.

[11] If the worker does not have access to complete insurance markets, the implications of the intertemporal labor supply model are unchanged, provided that the marginal utility of income is held constant. Since changes in productivity may be correlated with changes in the marginal utility of wealth, estimation of μ requires additional structure. (See our other paper.)

narginal utility of wealth follows a deterministic trend and contributes only the onstant $(\rho - r)$ to earnings and hours hanges.

In the labor supply model, the equation for the change in hours is identical to (10b). The equation for the change in earnings, however, is

$$\text{(11)}\quad \Delta \log g_{it} = (1+\eta)\Delta \log \theta_{it} + \delta(\rho - r) + \Delta b_{it} + \Delta u_{it}^{*}.$$

Equations (10a) and (11) are very similar. The statistically identifiable difference between the contracting and labor supply models arises from the different coefficients on the change in individual productivity. To clarify this point, we complete the models by specifying the covariance structures of individual productivity, preference variation, and measurement error.

We adopt a linear specification for individual productivity consisting of a permanent individual effect (θ_i), an aggregate time effect (d_t), a quadratic labor force experience effect, and a purely stochastic component (z_{it}):

$$\log \theta_{it} = \theta_i + d_t + \zeta_\theta x_{it} + \tfrac{1}{2}\xi_\theta x_{it}^2 + z_{it},$$

where x_{it} represents the labor force experience of individual i at the beginning of year t. Since labor force experience increases by one unit each year, the change in the logarithm of individual productivity is

$$\text{(12)}\quad \Delta \log \theta_{it} = \kappa_{\theta t} + \xi_\theta x_{i0} + \Delta z_{it},$$

where x_{i0} represents the labor force experience of individual i at the beginning of the survey and $\kappa_{\theta t}$ is a time effect that incorporates the change in the aggregate productivity shock as well as the change in average labor force experience.[12]

In a similar fashion, we assume that the preference variations (a_{it} and b_{it}) contain permanent individual effects, aggregate time effects, quadratic experience effects, and stationary, serially uncorrelated random components:

$$a_{it} = a_i + a_t + \zeta_a x_{it} + \tfrac{1}{2}\xi_a x_{it}^2 + \varepsilon_{ait},$$

$$b_{it} = b_i + b_t + \zeta_b x_{it} + \tfrac{1}{2}\xi_b x_{it}^2 + \varepsilon_{bit}.$$

These specifications permit tastes for consumption and leisure to exhibit homogeneous curvature over the life cycle. The vector of transitory deviations from the life cycle profile of preferences $(\varepsilon_{ait}, \varepsilon_{bit})$ is assumed to be independent and identically distributed for all i and t with an unrestricted contemporaneous covariance matrix. The first differences of the preference variations can be written as

$$\text{(13a)}\quad \Delta a_{it} = \kappa_{at} + \xi_a x_{i0} + \Delta \varepsilon_{ait},$$

$$\text{(13b)}\quad \Delta b_{it} = \kappa_{bt} + \xi_b x_{i0} + \Delta \varepsilon_{bit},$$

where κ_{at} and κ_{bt} are composite time effects that incorporate changes in a_t and b_t as well as changes in average labor force experience.[13]

Finally, we assume that the vector of survey measurement errors (u_{it}^{*}, v_{it}^{*}) contains permanent and purely transitory errors:

$$u_{it}^{*} = u_i^{*} + \varepsilon_{uit}, \quad v_{it}^{*} = v_i^{*} + \varepsilon_{vit}.$$

The permanent measurement error components, represented by u_i^{*} and v_i^{*}, model systematic deviations of the survey instrument from the theoretically appropriate concepts. We assume that the vector of transitory errors $(\varepsilon_{uit}, \varepsilon_{vit})$ is independent and identically distributed with an unrestricted contemporaneous covariance matrix. The first differences of the measurement errors can be written as

$$\text{(14a)}\quad \Delta u_{it}^{*} = \Delta \varepsilon_{uit},$$

$$\text{(14b)}\quad \Delta v_{it}^{*} = \Delta \varepsilon_{vit}.$$

[12] The effect $\kappa_{\theta t} = \Delta d_t + \zeta_\theta - \frac{1}{2}\xi_\theta + \xi_\theta t$, since $x_{it} - x_{it-1} = 1$ and $x_{it}^2 - x_{it-1}^2 = 2x_{i0} + 2t - 1$, where $x_{i0} =$ labor force experience at the beginning of the first survey period.

[13] The term $\kappa_{at} = \Delta a_t + \zeta_a + \xi_a t - \frac{1}{2}\xi_a$ and similarly for κ_{bt}.

Equations (14a) and (14b) indicate that only the transitory measurement errors contribute to the covariance structure of earnings and hours changes. Permanent response biases are eliminated by differencing.

Under the assumptions we have made, preference variation and survey measurement errors are statistically indistinguishable, since the first differences of both components represent first differences of uncorrelated vectors. For notational simplicity, we combine the transitory preference variation components, $\Delta\varepsilon_{ait}$ and $\Delta\varepsilon_{bit}$, with the transitory survey measurement error components, $\Delta\varepsilon_{uit}$ and $\Delta\varepsilon_{vit}$, to form a single vector of variance components $(\Delta u_{it}, \Delta v_{it})$. In the labor contract model, the preference variation and measurement error components of variance in earnings and hours changes are given by

$$\Delta u_{it} = \Delta\varepsilon_{ait} + \Delta\varepsilon_{uit}, \tag{15a}$$

$$\Delta v_{it} = \Delta\varepsilon_{bit} + \Delta\varepsilon_{vit}. \tag{15b}$$

In the labor supply model, on the other hand, the preference variation and measurement error components of variance in earnings and hours are given by

$$\Delta u_{it} = \Delta\varepsilon_{bit} + \Delta\varepsilon_{uit}, \tag{15a$'$}$$

$$\Delta v_{it} = \Delta\varepsilon_{bit} + \Delta\varepsilon_{vit}. \tag{15b$'$}$$

In either case, the vector $(\Delta u_{it}, \Delta v_{it})$ is independently and identically distributed across individuals with an arbitrary contemporaneous covariance matrix and a known autocovariance structure. Specifically, the vector $(\Delta u_{it}, \Delta v_{it})$ is a bivariate first-order moving average process with first-order autocorrelations equal to $-\frac{1}{2}$.[14] This simple autocorrelation structure reflects the following observation: if y_t is serially uncorrelated with variance σ^2, then the variance of Δy_t is $2\sigma^2$, the covariance of Δy_t with Δy_{t-1} is $-\sigma^2$, and the covariance between Δy_t and Δy_s is zero for $|t-s|>1$.

Combining equations (12)–(15), the equations for the changes in log earnings and log hours in the labor contracting model can be simplified to

$$\Delta \log g_{it} = \kappa_{gt} + \xi_g x_{i0} + \phi\Delta z_{it} + \Delta u_{it} \tag{16a}$$

$$\Delta \log h_{it} = \kappa_{ht} + \xi_h x_{i0} + \eta\Delta z_{it} + \Delta v_{it}, \tag{16b}$$

where κ_{gt} and κ_{ht} combine the aggregate time effects of equations (12) and (13); ξ_g and ξ_h combine the linear labor force experience effects of equations (12) and (13); Δu_{it} and Δv_{it} represent the combined preference variation and survey measurement errors from equation (15); and Δz_{it} represents the individual productivity variation from equation (12). For the labor supply model, the hours equations is the same as (16b). The equation for earnings, on the other hand, becomes

$$\Delta \log g_{it} = \kappa'_{gt} + \xi'_g x_{i0} + (1+\eta)\Delta z_{it} + \Delta u_{it}, \tag{17}$$

where κ'_{gt} combines the aggregate time effects of equations (12) and (13), and ξ'_g combines the linear labor force experience effects of changes in productivity and preferences. In general, the year effects κ_{gt} and κ'_{gt} and the experience slopes ξ_g and ξ'_g are different in (16a) and (17).

Neither the labor supply model nor the labor contracting model, however, imposes any restrictions on the year effects or experience slopes of equations (16) and (17). Under our assumptions, individual productivity changes and preference variations contribute three unrestricted time effects ($\kappa_{\theta t}$, κ_{at}, and κ_{bt}) to the changes in log earnings and log hours in the labor contract model, or two unrestricted time effects ($\kappa_{\theta t}$ and κ_{bt}) in the labor supply model. The cross-sectional means of $\Delta \log g_{it}$ and $\Delta \log h_{it}$ in period t (controlling for experience) are sufficient to estimate only two linear combinations of these effects (κ_{gt} and κ_{ht}). Similarly, there are three unrestricted labor force experience effects (ξ_θ, ξ_a, and ξ_b) in the labor contract model, or two unrestricted experience effects (ξ_θ and ξ_b) in the labor supply model. Again,

[14] This model also implies that the ratio of either first-order cross covariance $(\text{Cov}(\Delta \log g_{it}, \Delta \log h_{it+1})$, or $\text{Cov}(\Delta \log g_{it}, \Delta \log h_{it-1}))$ to the zero-order covariance $(\text{Cov}(\Delta \log g_{it}, \Delta \log h_{it}))$ is $-\frac{1}{2}$.

however, we can only identify two experience slopes (ξ_g and ξ_h).[15] Therefore, the coefficients of the multivariate regression of individual i's changes in log earnings and log hours on time effects and initial labor force experience are unrestricted by either model.[16]

Equations (16) and (17) do, however, provide a simple two-factor model for the residuals from the regression of changes in individual earnings and hours on time effects and labor force experience. According to these equations, unpredicted changes in earnings and hours contain a time-stationary preference and measurement error component, with a known autocorrelation structure, and a productivity component, with an arbitrary autocorrelation structure.

The relative contribution of productivity changes to earnings and hours changes depends on the parameters ϕ and η. In the absence of direct information on the variance of individual productivity shocks, these parameters are not separately identifiable from the covariance structure of earnings and hours changes. The critical parameter μ, which represents the ratio of ϕ to η, is identifiable from the *relative* covariances of earnings and hours changes, however. In particular, μ is identifiable if changes in earnings and hours exhibit second-order or higher autocorrelation, since the preference variation and measurement error components are assumed to contribute only first-order autocorrelation. Alternatively, μ is identifiable if the first-order autocorrelations of earnings and hours changes are not identically equal to $-\frac{1}{2}$, since the preference variation and measurement error components are assumed to have first-order autocorrelations equal to $-\frac{1}{2}$. Lastly, μ is identifiable if the autocovariances and cross covariances of earnings and hours changes are not time stationary, since the preference variation and measurement error components are assumed to be stationary.

While μ is potentially identifiable from the covariance structure of earnings and hours changes, the separate variance contributions of changes in productivity, preference shifts, and measurement errors are not identifiable. All three components contribute to the variances and first-order autocovariances of changes in earnings and hours. It is impossible to determine their separate contributions, however, without further assumptions on either the serial correlation properties of the productivity shocks, or the correlations of the measurement errors and preference variations.

Table 1 displays the theoretical formulas for the autocovariances and cross covariances of earnings and hours changes implied by equations (16) and (17). The variables $\Delta \log \tilde{g}_{it}$ and $\Delta \log \tilde{h}_{it}$ are defined as the deviations of $\Delta \log g_{it}$ and $\Delta \log h_{it}$, respectively, from their conditional (regression-adjusted) means given t and x_{i0}. We refer to these variables as experience-adjusted changes in log earnings and log hours. The formulas are written in terms of the parameter μ so that they apply to either the contracting or labor supply model.[17] Table 1 shows how the covariance structure of earnings and hours changes depends on the covariance structure of individual productivity changes (Δz_{it}), the covariance structure of

[15] If one assumes that life cycle tastes for leisure are linear (rather than quadratic) in labor market experience, then the life cycle labor supply model implies that the ratio of the labor force experience coefficient of earnings changes (ξ_g) to the labor force experience coefficient of hours (ξ_h) equals $(1+\eta)/\eta$. Provided that tastes for leisure are linear in experience, then, the intertemporal substitution elasticity may be estimated directly from the covariances of earnings and hours changes with experience. This technique is equivalent to the instrumental variables estimation schemes used by MaCurdy and by Joseph Altonji (1986). For the samples considered in this paper, the *PSID* yields an estimate of η equal to 1.52 (with a standard error of .44) and the *NLS* yields an estimate of η equal to -1.62 (with a standard error of .61) when this technique is used.

[16] The experience slopes of earnings and hours changes are actually restricted to be constant over time. Both the contracting model and the labor supply model imply this restriction, given our model for individual productivity and preference variations. The experience slopes from the *PSID* are consistent with this restriction ($\chi^2 = 25.80$ with 18 degrees of freedom, probability value $= .104$). The experience slopes from the *NLS* are not consistent with this restriction ($\chi^2 = 34.66$ with 8 degrees of freedom, probability value $\cong 0$).

[17] Expressing the covariances as functions of μ requires the definition of $\Delta \tilde{z}_{it} \equiv \eta \Delta z_{it}$.

TABLE 1—IMPLIED COVARIANCES OF EXPERIENCE-ADJUSTED CHANGES IN LOG EARNINGS AND LOG HOURS: CONTRACTING AND LABOR SUPPLY MODELS

Covariance Element	Implied Formulae
1) $\mathrm{Var}(\Delta \log \tilde{g}_{it})$	$\mu^2 \mathrm{Var}(\Delta \tilde{z}_{it}) + 2\sigma_u^2$
2) $\mathrm{Var}(\Delta \log \tilde{h}_{it})$	$\mathrm{Var}(\Delta \tilde{z}_{it}) + 2\sigma_v^2$
3) $\mathrm{Cov}(\Delta \log \tilde{g}_{it}, \Delta \log \tilde{h}_{it})$	$\mu \mathrm{Var}(\Delta \tilde{z}_{it}) + 2\rho_{uv}\sigma_u\sigma_v$
4) $\mathrm{Cov}(\Delta \log \tilde{g}_{it}, \Delta \log \tilde{g}_{it-1})$	$\mu^2 \mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-1}) - \sigma_u^2$
5) $\mathrm{Cov}(\Delta \log \tilde{h}_{it}, \Delta \log \tilde{h}_{it-1})$	$\mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-1}) - \sigma_v^2$
6) $\mathrm{Cov}(\Delta \log \tilde{g}_{it}, \Delta \log \tilde{h}_{it-1})$, $\mathrm{Cov}(\Delta \log \tilde{h}_{it}, \Delta \log \tilde{g}_{it-1})$	$\mu \mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-1}) - \rho_{uv}\sigma_u\sigma_v$
7) $\mathrm{Cov}(\Delta \log \tilde{g}_{it}, \Delta \log \tilde{g}_{it-2})$	$\mu^2 \mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-2})$
8) $\mathrm{Cov}(\Delta \log \tilde{h}_{it}, \Delta \log \tilde{h}_{it-2})$	$\mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-2})$
9) $\mathrm{Cov}(\Delta \log \tilde{g}_{it}, \Delta \log \tilde{h}_{it-2})$, $\mathrm{Cov}(\Delta \log \tilde{h}_{it}, \Delta \log \tilde{g}_{it-2})$	$\mu \mathrm{Cov}(\Delta \tilde{z}_{it}, \Delta \tilde{z}_{it-2})$

Note: $\Delta \log \tilde{g}_{it} \equiv \Delta \log g_{it} - \kappa_{gt} - \xi_g x_{i0}$, $\Delta \log \tilde{h}_{it} \equiv \Delta \log h_{it} - \kappa_{ht} - \xi_h x_{i0}$, $\Delta \tilde{z}_{it} \equiv \eta \Delta z_{it}$, $\rho_{uv} \equiv$ Correlation $(\Delta u_{it}, \Delta v_{it})$, $\sigma_u^2 \equiv \mathrm{Var}(\Delta u_{it})$, $\sigma_v^2 \equiv \mathrm{Var}(\Delta v_{it})$,

the preference variation and measurement error process (Δu_{it} and Δv_{it}), and μ. The formulas in Table 1 form the basis for our empirical test of the contracting model vs. the labor supply model.

III. A Test of the Contracting Model vs. the Labor Supply Model, using Longitudinal Data on Adult Males

The longitudinal earnings and hours data used in this paper are drawn from the *Panel Study of Income Dynamics* and the *National Longitudinal Survey of Men 45–59*. These surveys use substantially different methods for determining annual earnings and annual hours worked. The *PSID* collects information on various components of labor earnings while the *NLS* collects data on wage and salary income in a single question. The *PSID* likewise collects information on hours for both primary and secondary jobs. The *NLS*, on the other hand, collects annual hours information for the main job only. Our Data Appendix describes the variables we actually used and the survey questions from which these variables were derived.[18]

From the *PSID*, we selected 1448 male household heads whose records indicate nonzero earnings and hours in each year from 1969 to 1979 (the third through thirteenth waves of the survey). We included only those male household heads who were between the ages of 21 and 64 in all eleven sample years. The "one-employer" subsample was defined on the basis of answers to the questions about present employment status and reason for changing employment status. If an individual was currently employed, or temporarily laid off and reported having his present job for at least one year (including promotions), then the individual was considered to have the same employer as in the previous year. An individual with the same employer as in the previous year for all years from 1970 to 1979 was included in the one-employer subsample. There were 618 individuals who satisfied this condition. The remaining 830 individuals were included in the "multiple-employers" subsample. Every member of the multiple-employers subsample experienced at least one change of employer during the period from 1969 to 1979. Table 2 presents means and standard deviations of the changes in log real annual earnings, and log annual hours, as well as basic demographic variables for the *PSID* sample and subsamples.

From the *National Longitudinal Survey of Men 45–59*, we selected 1309 men whose

[18] See Survey Research Center (1981) and Center for Human Resources Research (1977, 1980) for documentation of the survey variables and procedures.

TABLE 2—SAMPLE AND SUBSAMPLE CHARACTERISTICS FOR THE *PSID* AND THE *NLS* OF OLDER MEN: MEANS AND STANDARD DEVIATIONS FOR THE INDICATED YEARS

	PSID				*NLS*[a]		
Year	All	One Employer	Multiple Employers	Year	All	One Employer	Multiple Employers
Change in Log Real Earnings[b]							
1969–70	2.5	1.8	3.0	1966–67	4.4	2.6	6.7
	(40.)	(24.)	(49.)		(31.)	(20.)	(41.)
1970–71	3.0	2.5	3.4	1967–69	4.6	4.0	5.3
	(40.)	(24.)	(48.)		(31.)	(23.)	(39.)
1971–72	6.9	5.8	7.7	1969–71	2.5	3.2	1.6
	(41.)	(30.)	(48.)		(31.)	(23.)	(38.)
1972–73	4.7	3.5	5.6	1971–73	−1.2	1.2	−4.4
	(37.)	(29.)	(41.)		(38.)	(29.)	(47.)
1973–74	−5.5	−3.8	−6.8	1973–75	−16.4	−6.9	−28.7
	(36.)	(26.)	(42.)		(54.)	(32.)	(72.)
1974–75	−4.2	−2.8	−5.3				
	(43.)	(28.)	(51.)				
1975–76	4.1	2.7	5.2				
	(47.)	(30.)	(57.)				
1976–77	2.5	2.8	2.3				
	(44.)	(25.)	(54.)				
1977–78	0.2	0.0	0.4				
	(44.)	(27.)	(53.)				
1978–79	−5.5	−4.4	−6.4				
	(42.)	(28.)	(51.)				
Change in Log Annual Hours[b]							
1969–70	−0.8	−1.8	−0.1	1966–67	−0.1	−0.8	0.8
	(35.)	(21.)	(42.)		(29.)	(19.)	(37.)
1970–71	−0.3	−0.6	−0.0	1967–69	−0.3	−0.5	−0.2
	(34.)	(18.)	(42.)		(27.)	(21.)	(33.)
1971–72	2.0	−0.1	3.6	1969–71	−0.9	−0.3	−1.6
	(34.)	(19.)	(42.)		(26.)	(19.)	(33.)
1972–73	1.9	2.1	1.7	1971–73	−1.2	0.9	−3.9
	(28.)	(19.)	(34.)		(29.)	(18.)	(38.)
1973–74	−4.1	−2.5	−5.3	1973–75	−10.9	−3.0	−21.1
	(27.)	(18.)	(32.)		(46.)	(22.)	(64.)
1974–75	−2.4	−1.1	−3.4				
	(33.)	(20.)	(41.)				
1975–76	0.6	−0.2	1.2				
	(38.)	(22.)	(47.)				
1976–77	0.3	0.6	0.1				
	(38.)	(20.)	(47.)				
1977–78	0.5	−0.9	1.6				
	(37.)	(21.)	(45.)				
1978–79	−4.2	−1.3	−6.4				
	(37.)	(24.)	(44.)				
Demographic Characteristics							
Age	35.8	38.2	34.1		49.1	48.9	49.5
	(9.)	(8.)	(9.)		(3.)	(3.)	(3.)
Potential Experience	18.9	21.4	17.1		34.4	33.8	35.2
	(11.)	(10.)	(11.)		(6.)	(5.)	(5.)
Percent Nonwhite	27.3	27.7	27.0		29.6	30.8	28.2
Sample Size	1448[c]	618	830		1309[d]	735	574

Note: Standard deviations are shown in parentheses.

[a] Statistics from the *NLS* are *not* at annual rates.

[b] Means and standard deviations times 100.

[c] Eight outliers with average hourly earnings greater than $100/hour (1967 dollars) have been deleted.

[d] Nine outliers with absolute changes in log earnings or log hours in excess of 3.5 have been deleted.

records indicate nonzero earnings and hours for each of the years 1966, 1967, 1969, 1971, 1973, and 1975.[19] We included only those males who were between the ages of 45 and 64 in all six sample years. The one-employer subsample was defined on the basis of the number of years the individual had worked for his current employer in 1971, and whether or not the individual worked for a different employer in 1973 or 1975. An individual who had worked for his current employer at least five years in 1971 and who did not change employers in either 1973 or 1975 was included in the one-employer subsample. There were 735 individuals who satisfied this condition. The remaining 574 individuals were included in the multiple-employers subsample. Means and standard deviations for the *NLS* sample are also presented in Table 2. For the interpretation of the *NLS* data, it is important to note that later waves of the survey were administered biennially. The changes in earnings and hours from 1969 to 1975 for the *NLS* therefore refer to changes in annual totals differenced over two-year intervals. These changes are *not* reported at annual rates in Table 2.

Table 2 shows that average age and potential labor force experience (age minus years of education minus five) differ substantially between the *PSID* and *NLS* samples because of design differences in the underlying surveys. The *NLS* workers are an average of 13.3 years older and 15.5 years more experienced than the *PSID* workers. We use experience-adjusted changes in earnings and hours in all subsequent calculations in this paper to correct for any systematic differences between the *PSID* and *NLS* samples arising from this difference in labor force experience. Both surveys also oversampled nonwhites. The percentage of nonwhites is similar in our two samples, however, and we make no further adjustments to account for the small difference in racial composition between them.[20] There are no important differences in age, labor force experience, or percentage nonwhite between the one-employer and multiple-employers subsamples of either survey.

Table 2 reveals three striking features of the individual earnings and hours changes from our two samples. First, the overall pattern of changes in earnings and hours is similar in both surveys. This conclusion applies when comparing all individuals, individuals with one employer, and individuals with multiple employers. The older sample (*NLS*) experienced slightly larger changes in earnings and hours during the 1973 to 1975 period than the younger sample (*PSID*). Individuals in the multiple-employer subsample of each survey also experienced substantially larger earnings and hours changes during this period than those in the one-employer subsamples. Second, there is significant nonstationarity in the cross-sectional variation of changes in earnings and hours. In the *PSID* sample, earnings and hours changes are most variable in the 1975–76 period and least variable in the 1972–73 period. In the *NLS* sample, on the other hand, these changes are most variable in the 1973–75 period and least variable in the 1969–71 period. Nonstationarity is equally apparent in the one-employer and multiple-employers subsamples of both surveys. Finally, earnings and hours changes are much less variable for the one-employer subsamples of both surveys.[21]

[19] These six survey years were the only waves in which comparable earnings and hours data were collected.

[20] Our *PSID* sample includes the *Survey of Economic Opportunity* subsample, which oversampled low-income households. Our *NLS* sample includes the black enumeration districts, which also oversampled low-income households.

[21] One might ask if variability of earnings and hours changes for the multiple-employers subsamples is similar to the one-employer subsamples if we consider only those years that do not involve employer changes. The answer is yes. In the *PSID* sample, in which we have annual data sampled at an annual rate, individuals experience substantial variability in earnings and hours changes during the three-year period surrounding the change in employer. For the *NLS* sample, in which we cannot perform such a detailed year-to-year analysis because we have annual data sampled at a biennial rate, it is still true that most of the difference in variation between the multiple-employers and one-employer subsamples occurs because of the variability contributed by the period in which the employer change actually occurred. Put differently, most of the added variability in earnings and hours changes for the multiple-employers subsample occurs around the time of employer changes.

TABLE 3—STATIONARY CROSS-COVARIANCE STRUCTURE FOR *PSID* AND *NLS* SAMPLES AND SUBSAMPLES[a]

Sample Cross-Covariance	*PSID*			*NLS*[b]		
	All	One Employer	Multiple Employers	All	One Employer	Multiple Employers
Earnings Autocovariances						
1) Var$[\Delta \log \tilde{g}_t]$	.172	.074	.245	.158	.074	.259
	(.011)	(.006)	(.017)	(.011)	(.010)	(.020)
2) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{g}_{t-1}]$	−.060	−.031	−.081	−.043	−.028	−.069
	(.006)	(.004)	(.009)	(.006)	(.006)	(.011)
3) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{g}_{t-2}]$	−.007	−.002	−.010	−.001	.002	−.004
	(.003)	(.002)	(.004)	(.003)	(.002)	(.008)
Hours Autocovariances						
4) Var$[\Delta \log \tilde{h}_t]$	.117	.040	.174	.108	.039	.191
	(.007)	(.003)	(.012)	(.010)	(.005)	(.020)
5) Cov$[\Delta \log \tilde{h}_t, \Delta \log \tilde{h}_{t-1}]$	−.035	−.016	−.050	−.038	−.018	−.066
	(.003)	(.002)	(.006)	(.006)	(.003)	(.014)
6) Cov$[\Delta \log \tilde{h}_t, \Delta \log \tilde{h}_{t-2}]$	−.011	−.000	−.019	.008	.001	.017
	(.002)	(.001)	(.003)	(.005)	(.003)	(.011)
Earnings/Hours Cross-Covariances						
7) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{h}_{t+2}]$	−.006	−.001	−.010	.001	.000	.002
	(.002)	(.001)	(.004)	(.004)	(.001)	(.009)
8) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{h}_{t+1}]$	−.023	−.005	−.037	−.015	−.002	−.033
	(.004)	(.001)	(.007)	(.004)	(.001)	(.010)
9) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{h}_t]$	.073	.011	.119	.063	.008	.126
	(.007)	(.001)	(.012)	(.007)	(.002)	(.014)
10) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{h}_{t-1}]$	−.020	−.003	−.033	−.010	−.002	−.022
	(.004)	(.001)	(.007)	(.004)	(.001)	(.009)
11) Cov$[\Delta \log \tilde{g}_t, \Delta \log \tilde{h}_{t-2}]$	−.002	.001	−.004	.007	.000	.015
	(.003)	(.001)	(.005)	(.005)	(.002)	(.010)
12) Goodness of Fit for Nonstationary *MA*(2)[c]	137.19	168.09	153.54	15.15	11.84	16.09
	(.053)	(.000)	(.006)	(.233)	(.459)	(.187)
13) Goodness of Fit for Stationary *MA*(2)[d]	180.74	168.18	175.03	79.11	33.21	72.33
	(.000)	(.000)	(.000)	(.000)	(.043)	(.000)

[a] Covariance matrix and standard errors based on equally weighted minimum distance estimates of cross-covariances. The standard errors are shown in parentheses.

[b] *NLS* estimates based on the four changes with gaps of two years: 1967–69, 1969–71, 1971–73, 1973–75.

[c] χ^2 statistic for nonstationary bivariate *MA*(2) vs. arbitrary process for the bivariate cross-covariance function. The statistic has 112 degrees of freedom for the *PSID* sample and 12 degrees of freedom for the *NLS* sample. Probability values are shown in parentheses.

[d] χ^2 statistic for stationary bivariate *MA*(2) vs. nonstationary bivariate *MA*(2) process for cross-covariance function. The statistic has 87 degrees of freedom for the *PSID* sample and 21 degrees of freedom for the *NLS* sample. Probability values are shown in parentheses.

Our theoretical analysis of the contracting and labor supply models focuses on their implications for the autocovariances and cross-covariances of earnings and hours changes. In particular, the identifiability of the parameter that distinguishes the two models depends critically on observable characteristics of the covariance matrix of earnings and hours changes. Therefore, we study this matrix in detail. Table 3 presents the average cross-covariances of the *PSID* and *NLS* samples.[22] The similarity between

There is no substantial difference between the one-employer and multiple-employer subsamples when data from employer changes are excluded. See Altonji and Christina Paxson (1985) for a detailed comparison of hours variability between job changers and stayers in the *PSID*.

[22] The complete covariance matrix of earnings and hours changes for the *PSID* contains 210 unique elements; the complete covariance matrix for the *NLS*

the two samples is also evident in their covariance structure. Both samples and all the subsamples exhibit strong positive correlations between contemporaneous changes in earnings and hours, and strong negative autocorrelation in earnings and hours changes. This similarity is even more remarkable since the *PSID* data represent year-to-year changes, while the *NLS* data represent changes in annual data over two-year intervals.

In comparison to the first-order autocovariances of earnings and hours changes, the second-order autocovariances are relatively small, although nonzero in the *PSID* at least. The higher-order autocovariances in both data sets (not reported in Table 3) are generally small and mixed in sign. Row 12 of Table 3 contains the statistics for a test that the third- and higher-order autocovariances of earnings and hours changes are jointly equal to zero. This hypothesis is not rejected for any of the *NLS* samples or for the complete *PSID* sample. These samples are therefore consistent with a (nonstationary) bivariate second-order moving average (*MA*(2)) model of earnings and hours changes.[23] On the basis of the test statistics in row 12 there is some evidence of third- and higher-order serial covariation in the two subsamples of the *PSID*. These covariances are of trivial magnitude, however, and we choose to assume that they are zero in the interest of parameteric simplicity.[24]

For both complete samples and for all the subsamples except the one-employer subsample of the *NLS*, there is also strong evidence of nonstationarity in the covariances of earnings and hours changes. The goodness of fit statistics for a stationary model of the cross-covariances of earnings and hours (up to second order) are recorded in the last row of Table 3. Judging by these statistics, at least one of the variance components generating the changes in earnings and hours in the *PSID* and *NLS* surveys is nonstationary.

Table 3 also shows that the first-order autocorrelations of earnings and hours changes are negative and smaller than one-half in absolute value for both samples and all the subsamples.[25] Similarly, the ratios of the first-order cross-covariances of earnings and hours changes to their corresponding zero-order covariances are all negative and smaller than one-half in absolute value.[26] In the framework of our two-factor model, the fact that these autocorrelations are smaller than one-half in absolute value is evidence of a productivity component in earnings and hours. A pure measurement error model of the data implies that these autocorrelations are all exactly equal to $-\frac{1}{2}$.

To summarize the evidence in Table 3, the covariance structure of changes in earnings and hours is consistent with a second-order bivariate moving average model. Third- and higher-order autocovariances and cross-covariances are approximately zero in both the *PSID* and *NLS* surveys. In addition,

contains 55 unique elements. Estimates of these matrices (with standard errors) for the complete samples are contained in our other paper. The covariances reported in Table 3 represent simple averages of the covariances for each year of the *PSID* or *NLS* survey.

[23] By a nonstationary second-order moving average representation, we mean that $\mathrm{Cov}(\Delta \log g_{it}, \Delta \log g_{it-j}) = 0$, $\mathrm{Cov}(\Delta \log h_{it}, \Delta \log h_{it-j}) = 0$, $\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it+j}) = 0$, and $\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it-j}) = 0$, for all $j \geq 3$; and all other variances and covariances are unrestricted.

[24] One explanation for the higher-order serial correlation of the *PSID* data as compared to the *NLS* data is the fact that the *NLS* data are sampled biennially. If an *MA*(2) model is appropriate for year-to-year (*PSID*) changes, for example, then biennial (*NLS*) changes follow an *MA*(1) model.

[25] In the *PSID* sample, the first-order autocorrelations of earnings changes are −.35 (overall), −.42 (one-employer), and −.33 (multiple-employers). In the *NLS* sample, the first-order autocorrelations of earnings changes are −.27 (overall), −.39 (one-employer), and −.27 (multiple-employers). Similarly, the *PSID* first-order autocorrelations of hours changes are −.30 (overall), −.40 (one-employer), and −.29 (multiple-employers). The *NLS* first-order autocorrelations of hours changes are −.35 (overall), −.46 (one-employer), and −.35 (multiple-employers).

[26] In the *PSID* sample the ratios $\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it+1})/\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it})$ and $\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it-1})/\mathrm{Cov}(\Delta \log g_{it}, \Delta \log h_{it})$ are −.32 and −.27 (overall), −.46 and −.27 (one-employer), and −.31 and −.28 (multiple-employers), respectively. In the *NLS* sample these ratios are −.24 and −.16 (overall), −.25 and −.25 (one-employer), and −.26 and −.18 (multiple-employers), respectively.

both samples and all the subsamples exhibit (*i*) second-order serial correlation, (*ii*) covariance nonstationarity, and (*iii*) first-order autocorrelations of earnings and hours changes less than one-half in absolute value. Since any one of these three conditions is sufficient to identify the relative contribution of productivity changes to earnings as compared to hours in our two-factor variance components model, the parameter μ is empirically identified.

Table 1 describes the expected values of the variances, autocovariances, and cross-covariances of experience-adjusted earnings and hours changes in terms of the autocovariance structure of individual productivity and the covariance structure of preference variation and measurement error. Estimation of μ and tests of the goodness of fit of the statistical model described in Table 1 require that we parameterize the autocovariance structure of individual productivity changes. We use two different parameterizations. In the first case, we assume that Δz_{it} is a stationary second-order moving average.[27] In the second case, we assume that Δz_{it} is a nonstationary second-order moving average.[28] If individual productivity is stationary, the bivariate process for earnings and hours changes described in Table 1 is stationary. While we have strong evidence against a stationary covariance structure, the advantage of a stationary model is that the sufficient statistics for estimation of the structural parameters are just the average variances and covariances reported in Table 3. If individual productivity is a nonstationary second-order moving average, on the other hand, the sufficient statistics for estimation of the structural parameters are all the elements of the complete covariance matrix of earnings and hours changes up to second order. In both cases, we use a method of moments estimator based on minimizing the distance between the sample covariance matrix and the theoretical covariance matrix implied by Table 1 to estimate μ and the goodness of fit of the structural models.[29]

Table 4 reports the goodness of fit statistics and the associated estimates of μ for the various samples and subsamples of the *PSID* and *NLS* data.[30] Panel A contains the estimates for a stationary parameterization of individual productivity changes. The goodness of fit statistics are large, even in comparison to the goodness of fit statistics for an unrestricted stationary covariance model (reported in row 13, Table 3). The estimates of μ are all bigger than one, and are actually larger for the one-employer subsamples than for the multiple-employer subsamples or the overall samples. The estimates of μ for the one-employer samples are relatively imprecise, however, and one is within two standard errors of both estimates.

Panel B of Table 4 contains the estimates of μ and the goodness of fit statistics for a nonstationary parameterization of individual productivity changes. This model fits the data better in all cases, although the estimates of μ are not much affected. The nonstationary model actually provides an acceptable fit to the one-employer subsample of the *NLS*. For the other subsamples and the two complete samples, the two-factor model of the covariance structure of earnings and hours changes is rejected.

The point estimates of μ from the one-employer subsamples provide evidence against the contracting model of earnings and hours changes, and in favor of the intertemporal labor supply model. The associated estimates of the intertemporal substitution elasticity

[27] In the *PSID*, this results in the addition of three parameters for the productivity process. In the *NLS*, because of the irregular timing of the survey, this results in the addition of six parameters.

[28] In the *PSID*, this results in 27 parameters for the productivity process. In the *NLS*, this results in 12 parameters for the productivity process.

[29] See Gary Chamberlain (1984) for a discussion of the statistical theory of these estimators, and the comparison between these estimators and the maximum likelihood estimators. Our goodness of fit measures are derived in Whitney Newey (1985).

[30] Estimation of μ requires one arbitrary normalization of the variance parameters in Table 1. We set the correlation of Δu_{it} and Δv_{it} to 0. All statistics reported in Table 4, including the estimate of μ, are invariant to the choice of normalization.

TABLE 4—ESTIMATED RELATIVE CONTRIBUTION OF PRODUCTIVITY TO EARNINGS AND HOURS FOR *PSID* AND *NLS* SAMPLES AND SUBSAMPLES USING STATIONARY AND NONSTATIONARY SPECIFICATIONS

	PSID			*NLS*[b]		
Definition	All	One Employer	Multiple Employers	All	One Employer	Multiple Employers
A. Stationary Model						
1. Relative Contribution of Productivity to Variance of Change in Log Earnings (μ)	1.05 (.078)	1.54 (.439)	1.02 (.079)	1.56 (.174)	4.39 (2.19)	1.39 (.154)
2. Elasticity of Intertemporal Labor Supply (η)	19.96 (30.9)	1.85 (1.50)	51.81 (213.)	1.77 (.563)	.29 (.190)	2.53 (.980)
3. Goodness of Fit for Structural Model[a]	335.90 (.000)	229.64 (.000)	304.32 (.000)	145.54 (.000)	50.23 (.036)	144.72 (.000)
B. Nonstationary Model						
4. Relative Contribution of Productivity to Variance of Change in Log Earnings (μ)	1.14 (.091)	2.46 (.781)	1.16 (.089)	1.23 (.100)	4.10 (2.15)	1.14 (.080)
5. Elasticity of Intertemporal Labor Supply (η)	7.27 (4.81)	.68 (.365)	6.40 (3.64)	4.34 (1.89)	.32 (.220)	7.14 (4.08)
6. Goodness of Fit for Structural Model[b]	261.15 (.000)	142.13 (.000)	223.01 (.000)	127.02 (.000)	33.84 (.206)	97.80 (.000)

Note: The standard errors are shown in parentheses.

[a]χ^2 statistic for stationary structural model vs. nonstationary bivariate *MA*(2) model. The statistic has 92 degrees of freedom for the *PSID* sample and 34 degrees of freedom for the *NLS* sample. Probability values are shown in parentheses.

[b]χ^2 statistic for nonstationary model vs. nonstationary bivariate *MA*(2) model. The statistic has 68 degrees of freedom for the *PSID* sample and 28 degrees of freedom for the *NLS* sample. Probability values are shown in parentheses.

(η) are recorded in rows 2 and 5 of Table 4.[31] In the *PSID* one-employer subsample, the estimates of η from the stationary and nonstationary models are 1.84 and .68, respectively. These estimates are larger than the instrumental variables estimates reported by Joseph Altonji (1986) and Thomas MaCurdy (1981) for *PSID* males, although they are based on a very different methodology. In the *NLS* one-employer sample, the estimates of η are .29 for the stationary model and .32 for the nonstationary model. These estimates are comparable to other estimates based on individual longitudinal data.

While the results from the one-employer subsamples are relatively favorable to the labor supply interpretation of earnings and hours changes, the results from the multiple-employer subsamples and the overall samples reveal a major difficulty with this interpretation. In the contracting model, changes in employer represent changes between contracts. The contract model therefore offers a simple explanation for the greater variability of earnings and hours for those who change employers than those who do not. The labor supply model, on the other hand, predicts the same structure of earnings and hours changes within and across jobs. The labor supply model by itself does not explain the higher variation in earnings and hours changes for those who change jobs. The labor supply model also predicts the same relative effect of productivity changes on earnings and hours for job changers and stayers. The point estimates of μ for the multiple-employer subsamples, however, are very different from the estimates based on the one-employer subsamples. In both the *PSID* and *NLS* multiple-employer subsamples, μ is precisely estimated and close to, but greater than, one. The implied estimates of the inter-

[31] The estimates of η are obtained from the formula $\eta = 1/(\mu - 1)$. If μ is near one, η will be imprecisely estimated and the point estimate of η will fluctuate substantially with relatively small changes in the point estimate of μ.

temporal substitution elasticity are large and imprecise.

The estimates of μ for individuals who change employers suggest that productivity changes affect earnings and hours proportionately. In other words, for these individuals, hours vary at fixed wage rates. One potential explanation for this finding in the framework of a labor supply model is that individuals cannot fully insure productivity risks. In this case, our estimation strategy confounds changes in productivity and changes in the marginal utility of wealth. Since changes in the marginal utility of wealth influence earnings and hours proportionately in the labor supply model, μ is biased towards one if the component of variance that we attribute to productivity changes includes changes in the marginal utility of wealth. In our other work, however, we find that estimates of μ are unaffected by controlling for changes in the marginal utility of wealth. The evidence that changes in earnings and hours occur at constant wage rates is inconsistent with either labor supply models or the contracting models considered in this paper. Fixed wage contract models have been considered by Abowd and Orley Ashenfelter (1981), and applied in the macroeconomics literature by Stanley Fischer (1977) and John Taylor (1980), among others. Our results for the job changers suggest that these models may be useful in the empirical analysis of individual data as well.

Finally, Table 4 also reports parameter estimates for the complete *PSID* and *NLS* samples. It is clear from these estimates that the characteristics of the multiple-employer subsamples carry over to the complete samples. In the complete samples, changes in productivity have slightly larger effects on earnings than hours, although we cannot easily reject the hypothesis that productivity-induced changes in hours occur at fixed wage rates (i.e., $\mu = 1$).

IV. Conclusion

Our goal in this paper was to develop an empirical strategy for testing between contracting and labor supply models. Such a test must rely on the fundamental distinction between these models: in contracting models, earnings represent optimal consumption, whereas in labor supply models, earnings represent the product of wage rates and hours of work. We derive a testable contrast between the two models based on the relative variability of changes in earnings and changes in hours. If the contracting model is correct, earnings are less variable than hours with respect to changes in productivity. If the labor supply model is correct, the reverse is true.

In order to apply the test, we specify a complete model of earnings and hours variation, including productivity components and components due to changes in tastes and measurement errors. This statistical model is itself testable, providing a check on the ability of either theory to explain the covariance properties of earnings and hours changes in longitudinal data.

We apply the model to longitudinal data from the *PSID* and *NLS* surveys. Generally speaking, the data are inconsistent with the simple covariance structure implied by either the labor supply or contracting model. Contrary to the implications of the contracting model, the contribution of productivity shocks to earnings is at least as large as the contribution to hours. This is true for individuals with the same employer over the entire period of the *PSID* and *NLS* surveys and more generally. From the point of view of the labor supply model, however, the implied intertemporal substitution elasticities are large and imprecise. A simpler interpretation of the data is that productivity-related changes in hours occur at fixed wage rates. We conclude that the specification and testing of fixed wage models for individual earnings and hours data should be a high priority for future research.

DATA APPENDIX

For the *Panel Study of Income Dynamics* we used an extract from the thirteen-year merged individual tape distributed through the Inter-University Consortium for Political and Social Research and documented by the Survey Research Center of the Institute for Social Research (1981 and previous years). Our sample consisted of all males on the thirteen-year merged individual tape with complete age and schooling data who were continuously heads of household from wave III to wave

XIII of the survey and who reported nonzero annual labor earnings and annual hours in each of the 11 waves. We included individuals from both the *Survey of Economic Opportunity* subsample and the Survey Research Center national probability subsample.

The following is a description of the *PSID* variables used. Numbers like Vxxxx refer to the variable numbers in the Survey Research Center codebooks for the thirteen-year merged individual tape. Survey questions are referenced by the question number and the exact question from the questionnaire.

ANNUAL EARNINGS: The variables used were: V1196, V1897, V2498, V3051, V3463, V3863, V5031, V5627, V6174, V6767, and V7413. These correspond to Survey Research Center's computed values for the head of household's total labor income in the calendar year before the survey. Annual earnings are computed from questions that changed somewhat from year to year. For 1979 (wave XIII) earnings are based on the sum of the answers to the following survey questions:

(K8) How much did you (HEAD) receive from wage and salaries in 1979, that is, before anything was deducted for taxes and other things?

(K9) In addition to this did you (HEAD) have any income from bonuses, overtime, or commissions?

(K10) How much?

(K11) Did you (HEAD) receive any other income in 1979 from professional practice or trade?

(K12a) How much from professional practice?

(K12b) How much from farming or market-gardening?

(K12c) How much from roomers or boarders?

Farmers and others with business income also answer a battery of questions on net farm income and total business income. Only the labor part of farm, business, and roomer income is added to variable V7413. The determination of the labor part of these variables is part of the coding process at the Survey Research Center.

ANNUAL HOURS: The variables used were V1138, V1839, V2439, V3027, V3423, V3823, V4332, V5232, V5731, V6336, and V6934. These correspond to Survey Research Center's computed value for the head of household's annual hours worked in the calendar year before the survey. The actual survey questions on which this variable was based changed from year to year. For 1979 (wave XIII) the questions were:

(C26) How many weeks did you actually work on your main job in 1979?

(C27) And, on average, how many hours a week did you work on your main job in 1979?

(C28) Did you work any overtime which isn't included in that?

(C29) How many hours did that overtime amount to in 1979?

(C42) How many weeks did you work on your extra jobs in 1979?

(C43) On the average, how many hours a week did you work on your extra jobs?

CHANGE OF EMPLOYER: Individuals were considered to have changed employer between the past calendar year and the survey year if they either (a) reported being unemployed at the time of the survey (approximately March of the survey years), or (b) reported being employed or temporarily laid off at the time of the survey but reported a change in jobs not associated with a promotion. Current employment status was measured from variables V1278, V1983, V2581, V3114, V3528, V3967, V4458, V5373, V5872, V6492, and V7095. This variable is the filter question that determines whether the head of household is asked the battery of questions about current employment (if he is employed or temporarily laid off) or unemployment (if he is unemployed). For individuals who were employed or temporarily laid off at the time of the survey, a change of employer was measured by variables V1282, V1988, V2586, V3119, V3534, V3986, V4490, V5399, V5890, V6501, and V7104. This variable is based on the answer to the question about what happened to the head of household's previous job. In wave XIII this question was: "(C12) What happened to the job you had before—did the company go out to business, were you laid off, promoted, were you not working, or what?". Promotions were not counted as changes of employer. All other answers were counted as a change of employer.

SEX: The variables used were V7492, V7509, V7526, V7547, V7561, V7576, V7601, V7653, V7687, and V7714. Only males were included.

HEAD OF HOUSEHOLD: The variables used were V7490, V7507, V7524, V7545, V7559, V7574, V7599, V7624, V7651, V7685, and V7712. Only heads of household for all eleven waves from wave III to wave XIII were included.

SCHOOLING: The variables used were V0313, V0794, V1485, V2197, V2838, V3241, V3663, V4198, V5074, V5647, V6194, V6787, and V7433. Maximum completed schooling was determined by examining all schooling variables. This was the schooling level used.

AGE: The variables used were V7460, V7476, and V7491. Age was determined for the first year an individual entered the sample, then adjusted to age in 1969.

For the *National Longitudinal Survey of Men 45–59, 1966 to 1975*, we used an extract from the public use data file release 75A distributed through the Inter-University Consortium for Political and Social Research and documented by the Center for Human Resource Research (1977, 1980). We used data from survey years 1966, 1967, 1969, 1971, 1973, and 1975. Annual data from the survey year 1966 refer to calendar year 1965. Subsequent annual data refer to the twelve months preceding the actual interview—approximately June 1966 to May 1967 for the 1967 survey and approximately July of the previous year to June of the survey year for the subsequent surveys. Our sample consisted of all males who had valid age and schooling data and reported nonzero annual earnings and annual hours for the years we studied.

The following is a description of the *NLS* variables used. Numbers like Vxxxx refer to the Center for Human Resource Research codebook variables numbers (not the reference numbers) for the release 75A public use tape. (Some variables are assigned two consecutive variable numbers.) Survey questions are referenced by the question number but only the facimile question in the public use codebooks is reproduced.

ANNUAL EARNINGS: The variables used were V0263–4, V0784–5, V1280–1, V3166–7, V2528–9, V2685–6. For the first two survey years, these variables represent the answer to the question: "(63A) What was your income from wages and salary in 1965?" (Example from 1966 survey). In the subsequent years these variables represent the answer to the question: "(16) What was your income from wages and salary in the past year?" (Example from 1975 survey).

ANNUAL HOURS: The variables used were hours per week: V0082, V0660, V1128, V1581, V2520, V2675, and weeks per year: V0589, V1022, V1168, (V2421 with V2461), V2519, V2674. There is substantial survey-to-survey variation in the questions used to measure these hours concepts. In 1966 the hours per week question was: "(11B) What were the usual number of hours per week worked in 1965?". In 1967 the question was: "(7B) What is the number of hours worked at your current or last job?". In 1971 the question referred to the current job only. In 1973 and 1975 the question was "(12B) What is the number of hours per week usually worked during the weeks worked in the past year?" (Example from 1975 survey). The weeks worked per year variable is a Center for Human Resource Research recode of the raw data for the survey years 1966 to 1971. For the 1971 survey we recoded the weeks worked variable, which refers to weeks worked since the last interview, into weeks worked in the last year by dividing the number of weeks worked since the last interview (V2421) by the number of weeks since the last interview (V2461) and multiplying by 52. In 1973 and 1975 the variable refers to the question: "(12A) What is the number of weeks worked in the past year?" (Example from 1975 survey).

EMPLOYER CHANGES: The variables used were V2406, V2548, and V2708. The first of these is reported "tenure at current job" in the 1971 survey, which is a recode of the question: "(6H) What is the year you started working at your current job?". The last two variables are a recode of the answer to the question: "(Check Item C) Is the date you started working at your current job September 1, 1971 or later?" (Example from 1973 survey). Individuals with 1971 tenure greater than five years and no reported change of employer in the 1973 and 1975 surveys were treated as having the same employer for all years.

AGE: The variable used was V0024, age in 1966.

EDUCATION: The variable used was V0611, highest grade completed.

REFERENCES

Abowd, John M. and Ashenfelter, Orley, "Anticipated Unemployment, Temporary Layoffs and Compensating Wage Differentials," in Sherwin Rosen, ed., *Studies in Labor Markets*, Chicago: University of Chicago Press, 1981, pp. 141–70.

______ **and Card, David,** "On the Covariance Structure of Earnings and Hours Changes," NBER Working Paper No. 1832, February 1986.

Abraham, Katharine and Farber, Henry, "Job Duration, Seniority, and Earnings," unpublished manuscript, MIT, December 1985.

Altonji, Joseph A., "Intertemporal Substitution in Labor Supply: Evidence from Micro Data," *Journal of Political Economy*, June 1986, *94*, S176–S215.

______ **and Paxson, Christina H.,** "Job Characteristics and Hours of Work," unpublished manuscript, Princeton University, June 1985.

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria." *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin N., "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Brown, James N., "How Close to an Auction is the Labor Market," *Research in Labor Economics*, 1982, vol. 5, 182–235.

Chamberlain, Gary, "Panel Data," in Zvi Griliches and Michael Intriligator, eds., *The Handbook of Econometrics*, Vol. 2, New York: North-Holland, 1984.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Policy Economy*, February 1977, *85*, 191–205.

Friedman, Milton, *A Theory of the Consumption Function*, NBER, Princeton: Princeton University Press, 1957.

Ghez, Gilbert and Becker, Gary S., *The Allocation of Time and Goods over the Life Cycle*, NBER, New York: Columbia University Press, 1975.

Gordon, Donald, "A Neo-Classical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, *12*, 431–59.

Hall, Robert E., "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91–123.

______, "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, September 1982, *72*, 716–24.

Hart, Oliver D., "Optimal Labor Contracts under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–35.

Heckman, James J., "Life Cycle Consumption and Labor Supply: An Explanation of the Relationship between Income and Consumption over the Life Cycle," *American Economic Review*, March 1974, *64*, 188–94.

______, "A Life-Cycle Model of Earnings, Learning, and Consumption," *Journal of Political Economy*, August 1976, *84*, S11–S44.

Killingsworth, Mark, *Labor Supply*. New York: Cambridge University Press, 1983.

Lucas, Robert E. Jr. and Rapping, Leonard A., "Real Wages, Employment and Inflation," *Journal of Political Economy*, September/October 1969, *77*, 721–54.

MaCurdy, Thomas E., "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, December 1981, *89*, 1059–85.

Modigliani, Franco and Brumberg, Richard, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," *Post Keynesian Economics*. New Brunswick: Rutgers University Press, 1954, 388–436.

Newey, Whitney K., "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, September 1985, *29*, 229–56.

Oi, Walter, "Labor as a Quasi-Fixed Factor," *Journal of Political Economy*, December 1962, *70*, 538–55.

Pencavel, John, "Labor Supply of Men: A Survey," in Orley Ashenfelter and Richard Layard, eds., *The Handbook of Labor Economics*, forthcoming 1987.

Rosen, Sherwin, "Short Run Employment Variation on Class-I Railroads in the U.S., 1947–1963," *Econometrica*, July/October 1968, *36*, 511–29.

______, "Implicit Contracts: A Survey," *Journal of Economic Literature*, September 1985, *23*, 1144–76.

Taylor, John, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

Topel, Robert and Welch, Finis, "Efficient Labor Contracts with Employment Risk," unpublished manuscript, University of Chicago, 1986.

Center for Human Resource Research, *National Longitudinal Survey of Men 45–59, 1966 to 1975: Codebook*, Columbus: Ohio State University, 1977.

______, *The National Longitudinal Surveys Handbook*, Columbus: Ohio State University, 1980.

Survey Research Center, *A Panel Study of Income Dynamics Procedures and Tape Codes 1980 Interviewing Year Wave XIII: A Supplement*, Ann Arbor: Institute for Social Research, University of Michigan, 1981 (and prior years).

Queues, Rations, and Market: Comparisons of Outcomes for the Poor and the Rich

By RAAJ KUMAR SAH*

This paper compares outcomes of alternative allocation systems (queues, convertible and nonconvertible rations, and unhindered market) to distribute limited quantity of a deficit good among heterogeneous individuals. It is shown that, for the poor, the ranking of systems (from better to worse) is convertible rations, nonconvertible rations, queues, and nonintervention. The rich are better off under nonintervention than under other systems. These and other positive results are robust to certain types of commodity taxes and administrative costs.

"Nonmarket" allocation systems such as rationing and queues are not only extensively employed in many less developed countries and centrally planned economies, but also their consequences are issues of important controversies. There is a wide range of features that such allocation systems exhibit; for instance, the rationed good is not convertible (i.e., individuals cannot exchange this good in secondary markets) in some rationing and queue systems, whereas it is partly or fully convertible in others.[1]

Each of the above allocation systems leads to a markedly different distribution of welfare among various individuals in the economy, and these welfare distributions are quite different, in turn, from the one that would emerge if the government were not intervening. The primary objective of this paper is to compare the welfare of specific groups of individuals (particularly the poor and the rich) when the limited supply of a good (the deficit good) is allocated through alternative allocation systems, including nonintervention. I do this in two steps: (*i*) I ascertain the utilities of various groups of individuals under each allocation system, and then (*ii*) I take each pair of allocation systems and attempt to determine whether a specific group of individuals is better off under one allocation system or another.

My analysis is *positive*, and it is not my objective here to determine the societal desirability of alternative allocation systems. I believe, however, that analyses of the kind developed in the present paper can contribute significantly to typical political or normative debates about whether, when, and how governments ought to intervene in markets. For instance, a main argument often given in favor of the queue or the ration system is that (since direct income subsidies to the poor are not feasible) these allocation systems might be effective ways of helping the poor. My comparisons of the welfare of the poor under alternative allocation systems can help to recognize some of the circumstances when such arguments are useful and when they are not.

The specific allocation systems which I compare here are: nonintervention, convertible and nonconvertible rations, and the queue system (without secondary trade).[2] I show that

(*i*) *For the poor, the ranking of allocation systems (from better to worse) is convertible rations, nonconvertible rations, the queue system, and nonintervention.* The queue system, thus, does not turn out to be rela-

*Yale University, New Haven, CT 06520. I thank Martin Weitzman and three anonymous referees for comments on an earlier version of this paper.

[1] These systems have been employed and debated in developed countries as well, particularly in the context of external hostilities.

[2] See a more detailed version of the present paper (1986) for positive comparisons of some other systems such as the queue system with secondary trade, and the bundling system (where the deficit good is bundled with some other good).

tively as beneficial to the poor as it is often thought to be. Also, governments frequently attempt to enforce nonconvertibility of rations. Such an emphasis is potentially harmful to the poor.

(*ii*) *The rich are better off under nonintervention than they are under other allocation systems. Also, the rich are better off under convertible rations than they are under the queue system.* These results, as we shall see, are understandable consequences of the high wages and large endowments that the rich typically have.

It is often believed that no one can be worse off, and some individuals must be better off, under convertible rations than under nonconvertible rations, because there are gains to trade in the former system. But this view is incorrect because, as James Tobin (1952) had rightly argued, the convertibility of rations may alter individuals' incentives to buy the rationed good. Consequently, convertible rations are not always weakly Pareto superior to nonconvertible rations. I demonstrate this important aspect of rationing.

A methodological aspect of this paper is that the standard tools of marginal analysis are not usable here because alternative allocation systems result in equilibria which cannot be assumed to be in the neighborhood of one another. Yet, as we shall see, my results are robust not only to many parameters of the economy but also to certain types of commodity taxes and administrative costs. An additional strength of my pairwise comparisons among alternative systems is that the comparison between any two systems does not depend on whether a third system is considered feasible or not. For instance, nonintervention may not be a realistic alternative in centrally planned economies. In these contexts, the relevant comparisons are those among alternative government managed systems (i.e., among the rationing systems and the queue system).

A central contribution to the comparison of allocation systems is by Martin Weitzman (1977, pp. 517–19) in which he compared, based on a normative criterion of "satisfying the needs of the population," the allocation of a fixed quantity of the deficit good through nonconvertible rations versus a "price system." My analysis is different in not only the scope (I compare several important allocation systems in addition to the two that he does) and the emphasis (mine is on obtaining positive results, whereas his is on normative analysis based on a specific social criterion), but also in a critical aspect of the model of the price system (discussed later).

This paper is not related to the important literature which has extended the theory of second-best to instruments such as rations and queues. For instance, Roger Guesnerie and Kevin Roberts (1984) show that, starting from a second-best situation, a government can do better under certain circumstances if nonconvertible rations are partly introduced into an economy. Sam Bucovetsky (1984) shows that the same is possible if a queue system is partly introduced into an economy. The underlying economic reason is simple: the government cannot do worse by having additional policy instruments (whatever the instruments might be, provided it is assumed that there are no administrative costs) and it may do strictly better under some circumstances, regardless of what the social criterion might be.

The present paper has a different aim. My motivation here is not to study rations or queues as *additional* (and costless) policy instruments through which the government can do better, based on some criterion. Instead, my motivation is to examine and compare rations, queues, and market as *alternative* allocation systems.[3] In Section I, I derive the expressions for individuals' utilities under alternative systems. The method for comparing an individual's utility is summarized in Section II. Alternative systems are then compared to one another in Section III.

[3] I do not consider mechanisms such as nonlinear pricing schemes (with arbitrary nonlinearities) because such schemes are not feasible for consumption goods. In fact, only simple allocation systems, such as those considered in this paper, are typically feasible because of reasons such as the unavailability of information, and the limitations on third-party enforceability. My forthcoming paper with Joseph Stiglitz discusses some of the sources and the consequences of the restrictions on policy instruments available in *LCD*s.

I. Individuals' Utilities under Alternative Allocation Systems

First, I determine the utility levels of different individuals under four allocation systems: nonintervention (market), nonconvertible rations, convertible rations, and the queue system. These systems are respectively denoted by superscripts $I = M$, R, C, and Q. Individuals are denoted by the superscript h. The variable n^h is the proportion of individuals of type h in the economy, $n^h > 0$, and $\Sigma_h n^h = 1$.

Denote the available supply (per capita) of the deficit good by X, and its unit cost by p. For individual h, x^h, and V^h, respectively, denote the demand function for the deficit good, and the indirect utility function. I assume that the market demand for the deficit good would exceed the available quantity (i.e., there would be a "shortage") if its market price were to be set equal to its unit cost.[4] That is,

$$\sum_h n^h x^h(p, m^h) > X, \tag{1}$$

where m^h is the (full) income of individual h if the market price of the deficit good is p.[5]

Under nonintervention, therefore, private firms (owners of the deficit good) adjust the consumer price of the deficit good to equate its demand and supply. Under a government-managed system, the government procures the available quantity of the deficit good at its unit cost p, and distributes it through one or another allocation system.[6] I assume at present that the price of the deficit good that the government charges at its shops is also p; issues concerning administrative costs and commodity taxes are discussed later.

For individual h, let x^{hI} and V^{hI} denote the quantity of the deficit good consumed, and the utility obtained, under the allocation system I. The economywide consumption of the deficit good equals its available quantity under each system; that is,

$$\sum_h n^h x^{hI} = X, \quad \text{for} \quad I = M, R, C, Q. \tag{2}$$

I now obtain the expressions for V^{hI} for various systems, which are needed for later comparisons.

A. *Nonintervention*

The individual h owns (through partial ownership of firms) $\alpha^h X$ units of the deficit good. Naturally, $\alpha^h \geq 0$, and $\Sigma_h n^h \alpha^h = 1$. If the market-clearing price is p^M, then the full income of individual h is $m^h + \alpha^h(p^M - p)X$.[7] Thus

$$V^{hM} = V^h\big(p^M, m^h + \alpha^h(p^M - p)X\big) \tag{3}$$

$$\text{and} \quad x^{hM} = x^h\big(p^M, m^h + \alpha^h(p^M - p)X\big).$$

The market price p^M is obtained by substituting the expression for x^{hM} into (2). We restrict our analysis to those situations where the aggregate market demand curve for the deficit good is downward sloping in its price. The relevant implication of this restriction, from (1) and (2), is that the market price p^M is higher than p. This implication is consistent with the intuition that systems such as rationing are typically employed in those situations where the market allocation would entail a significant rise in the price of the deficit good.

[4] In fact, it is under these conditions that governments typically intervene by employing allocation systems such as rations or queues.

[5] For notational convenience, an individual's wage rate and the prices of nondeficit goods are suppressed in the arguments of his demand function and his indirect (individualistic) utility function.

[6] In those contexts where nonintervention is not a feasible alternative (for instance, when the deficit good is produced in the public sector), p is the unit cost to the government.

[7] Where $\alpha^h(p^M - p)X$ is the profit from ownership which nonintervention brings to individual h. Weitzman's model of a price system assumes for simplicity that these profits disappear altogether. But as we shall see, these profits (no matter how they are distributed among individuals) play a critical role in determining not only the welfare and the consumption of individuals but also the market-clearing price.

B. *Nonconvertible Rations*

Under this system, individuals can buy (at government shops) up to a fixed quantity, X^R, of the deficit good, but no more, and resale is not permitted. Naturally, the population self-selects itself into two groups. The first group consists of those who wish to buy the deficit good in quantities smaller than or equal to X^R. These individuals are not constrained by rationing. For them,

$$(4) \qquad V^{hR} = V^h(p, m^h).$$

The second group consists of those who want to consume more deficit good than X^R, but are constrained to consume only X^R. A convenient representation of an individual's utility under a rationing constraint is to define the virtual price of the deficit good for person h to be p^{hR}, which is obtained from $x^h(p^{hR}, m^h + (p^{hR} - p)X^R) = X^R$. Then, this person's consumption behavior under rationing is the same as that in the hypothetical case when he faces price p^{hR}, receives an income transfer $(p^{hR} - p)X^R$, and faces no rationing. Therefore, the utility level of person h can be expressed as

$$(5) \quad V^{hR} = V^h(p^{hR}, m^h + (p^{hR} - p)X^R),$$

where $p^{hR} > p$.[8]

I assume that there are at least some individuals in the economy (the poorest persons are among them) who do not (or cannot) buy the maximum ration quantity X^R. This, I believe, is a more accurate representation in most situations (particularly in *LDC*s) than to assume that everyone buys the maximum ration quantity. It follows then that

$$(6) \qquad X^R > X.$$

C. *Convertible Rations*

If rations purchased from the government shops can be subsequently traded, and if the resulting equilibrium price of the deficit good is higher than p, then everyone would buy the full quantity of available ration. The ration per person is thus X. If p^C denotes the equilibrium price, then

$$(7) \quad V^{hC} = V^h(p^C, m^h + (p^C - p)X).$$

The price p^C is obtained by substituting $x^{hC} = x^h(p^C, m^h + (p^C - p)X)$ into (2). Comparison of (7) with (3) shows, as one might expect, that the key difference between nonintervention and convertible rations is that, in the latter system, the government intervention has effectively equalized the ownership of the deficit good. Since the income distribution in these two cases is different, p^C and p^M are not the same, in general. But $p^C > p$, given my earlier restriction that the aggregate demand curve for the deficit good is downward sloping in price.

D. *Queues*

The wage rate for individual h is denoted by w^h. I assume for brevity that the waiting time per unit purchase, t, is not significantly affected by the quantity purchased. This representation approximates those cases where individuals make several purchases within a single decision period; for instance, because the deficit good is dispensed in small lots, or because private storage of the good is expensive.[9] The opportunity price of the deficit good to individual h is $p + tw^h$, and his

[8] To see that $p^{hR} > p$, note from (5) that $\partial V^{hR}/\partial X^R = \mu^h(p^{hR} - p)$, where μ^h is the positive marginal utility of income for this person. Also, $\partial V^{hR}/\partial X^R$ is positive because this person wants to consume more of the deficit good. Hence, $p^{hR} > p$. See J. Peter Neary and Roberts (1980) for additional details of this representation.

[9] My analysis is readily extended, however, to a more general specification in which t differs across individuals and it is determined in part by individuals' decisions concerning the quantity and the frequency of their purchases. In fact, it can be easily verified that if t^h denotes the waiting time per unit purchase, then a sufficient condition under which the results I derive later remain unaffected is that the waiting cost per unit purchase, $t^h w^h$, is very small (but positive) at the lower end of the wage distribution, and that this cost is relatively large at the upper end of the wage distribution.

utility level is

$$(8)\qquad V^{hQ}=V^h(p+tw^h,m^h),$$

where t is determined from $x^{hQ}=x^h(p+tw^h,m^h)$ and (2).

I assume that the prices of the nondeficit goods (i.e., of goods other than the deficit good) and the wage rate of any given individual are not significantly different under the four allocation systems described above. This would be the case if, for example, the supply elasticities of the nondeficit goods and the demand elasticities for different types of labor are large.

II. Method for Comparing an Individual's Utility

If I and J represent two different allocation systems, then I want to ascertain whether the individual h is better off or worse off under I; that is, whether V^{hI} is larger or smaller than V^{hJ}. For notational brevity, let p^{hI} and m^{hI} denote the price of the deficit good and the income, corresponding to individual h, under the system I. Let p^{hJ} and m^{hJ} denote the respective variables under the system J. Then the individual is obviously better off under the system I if $m^{hI}\geq m^{hJ}$ and $p^{hI}\leq p^{hJ}$, with at least one strict inequality. This is because a higher income or a lower price (or both) yield a higher utility.

To deal with the remaining cases, in which one of the two allocation systems entails a higher price but also a higher income for an individual, define the metric

$$(9)\quad \Delta^h(I,J)=(m^{hI}-m^{hJ})+(p^{hJ}-p^{hI})x^{hJ}.$$

Then it can be shown that

$$(10)\quad V^{hI}>V^{hJ},\qquad \text{if}\quad \Delta^h(I,J)\geq 0.$$

A revealed preference argument underlying (10) is as follows. If $\Delta^h\geq 0$, then (9) implies that this individual could have purchased, in allocation system I, the same bundle of goods as he did in the allocation system J. The individual's actual purchase under the allocation system I, however, was different. Therefore, the individual h must be better off under I.[10]

Note that this method does not yield a verdict when the metric (9) is negative or when its sign cannot be ascertained based on the available information, but it is the best available method for comparing an individual's utility under two different situations, without restricting his preferences. In the analysis below, therefore, I compare as many pairs of allocation systems as are possible based on the above method.

III. Comparisons Among Alternative Allocation Systems

In this section, I compare the outcomes of the allocation systems described in Section I. I do this first for the poor, then for the rich. I then compare certain aspects of convertible vs. nonconvertible rations. Issues concerning commodity taxation and administrative costs are examined at the end.

A. *Comparisons for the Poor*

The poor are denoted by $h=1$. Since the poor belong to the lower tail of the distribution of incomes and wages, their demand for the deficit good under nonconvertible rations is smaller than the per capita available quantity. That is,

$$(11)\qquad x^{1R}<X.$$

No special assumption is needed for the poor to behave this way; the budget constraint itself will generate such a demand behavior at sufficiently low incomes. Also, the poor do not get any part of the profit under nonintervention; this is a reasonable assumption because the poor do not typically possess ownership of firms. That is, $\alpha^1=0$, and from (3): $V^{1M}=V^1(p^M,m^1)$. I now derive the following result: The ranking of allocation systems for the poor (from

[10] Expression (10) can also be established by using the standard concavity properties of expenditure functions. See my 1986 paper.

better to worse) is convertible rations, nonconvertible rations, the queue system, and nonintervention.

Begin by comparing convertible rations to nonconvertible rations. Expressions (4), (7), and (9) yield

$$\Delta^1(C,R) = (p^C - p)(X - x^{1R}). \tag{12}$$

Using (11) and recalling that $p^C > p$, it follows that (12) is positive. Therefore, the poor are better off under the ration system with convertibility than they are if rations are nonconvertible. The reason for this is as follows. Convertibility of rations brings an income gain to the poor, but it also entails a higher price for the deficit good. On the whole, the poor are better off with convertibility because the (income-producing) ration quantity they can get under this system exceeds the quantity of the deficit good they consume under nonconvertible rations.

The comparison between nonconvertible rations and the queue system is straightforward since, from (4) and (8), the poor have the same income under these two systems, but they face a higher price of the deficit good under the latter. This is because the queue system entails an *extra* cost of waiting, small though this extra cost may be for the poor. Thus, $V^{1R} > V^{1Q}$. Finally, compare $V^{1M} = V^1(p^M, m^1)$ to (8). The poor have the same income under the queue system and nonintervention, but the respective prices for the deficit good are $p + tw^1$ and p^M. Now recall that $p^M > p$. It follows then that a person with sufficiently low wage is better off under the queue system than under nonintervention.

B. *Comparisons for the Rich*

The rich are denoted by $h = r$, and they belong to the upper tail of the distribution of incomes and wages. As one would expect, the comparisons between nonintervention and other systems depend, in part, on the ownership of the deficit good that the rich have under nonintervention. I show here that: The rich are better off under nonintervention than under other allocation systems, if their ownership of the deficit good under nonintervention is large; specifically if

$$\alpha^r X \geq x^{rI}, \quad \text{for} \quad I = R, C, Q. \tag{13}$$

That is, if the rich own more deficit good under nonintervention than what they consume under other systems.

The condition (13) is automatically satisfied in a two-class economy because, in this case, the rich own all of the deficit good under nonintervention, but (regardless of the allocation system) the poor consume at least some of the deficit good. In fact, we expect the condition (13) to be satisfied in a multiclass economy as well, because the rich typically own proportions of firms' shares which are far in excess of the proportions of the outputs of firms that they consume.

To establish the above results, I obtain the following from (3), (5), (7), (8), and (9)

$$\Delta^r(M,R) = (p^M - p)(\alpha^r X - X^R) \tag{14}$$

$$\Delta^r(M,C) = (p^C - p)(x^{rC} - X) + (p^M - p)(\alpha^r X - x^{rC}) \tag{15}$$

$$\Delta^r(M,Q) = (p^M - p)(\alpha^r X - x^{rQ}) + tw^r x^{rQ}. \tag{16}$$

Recall that $p^M > p$, and $p^C > p$. Using (13), thus, (14) and (16) are nonnegative. Further, under convertible rations, the consumption of the deficit good by the rich would typically not be less than the economywide average consumption; that is $x^{rC} \geq X$.[11] Hence, (15) is also nonnegative.

We can also show that those with very high wages (which includes the rich) are better off under convertible rations than under the queue system. Specifically, expressions (7), (8), and (9) yield:

$$\Delta^h(C,Q) = (p^C - p)X + [tw^h - (p^C - p)]x^{hQ}.$$

[11] Sufficient conditions for this to be the case are that the deficit good is normal, and that the individuals' tastes are similar.

Since $p^C > p$, the preceding expression is positive if $w^h \geq (p^C - p)/t$.

C. *Convertible vs. Nonconvertible Rations*

To show that certain individuals are better off under nonconvertible rations than under convertible rations, I consider those whose consumption of the deficit good under convertible rations is between X and X^R; that is, $X^R \geq x^{hC} \geq X$. Among these individuals, there could be two types: those whose consumption is not constrained under nonconvertible rations, and those whose consumption is constrained. For the former type, expressions (4), (7), and (9) yield

$$(17) \qquad \Delta^h(R,C) = (p^C - p)(X^{hC} - X).$$

For the latter type, expressions (5), (7), and (9) yield

$$(18) \qquad \Delta^h(R,C) = (p^{hR} - p)(X^R - x^{hC}) + (p^C - p)(x^{hC} - X).$$

Both (17) and (18) are nonnegative because $p^C > p$, and $p^{hR} > p$. Thus, this entire group of individuals is better off under nonconvertible rations than under convertible rations.

The intuition behind this result can be seen in two steps. First, under convertible rations, everyone has an incentive to buy the maximum quantity of rations available; consequently, this quantity equals X. There is no corresponding incentive under nonconvertible rations. Therefore, the maximum ration quantity, X^R, is larger than X, because there are individuals who do not buy the maximum ration quantity. Second, recall that the convertibility of rations implies a higher price of the deficit good, but also an income gain $(p^C - p)X$. Thus, for those individuals whose consumption under convertible rations is larger than X but smaller than X^R, the loss due to higher price exceeds the income gain from convertibility.

Note that the above result is based on my assumption that some individuals in the economy do not (or cannot) buy the maximum ration quantity under the nonconvertible ration system. Under the less realistic assumption that everybody buys the maximum quantity under the nonconvertible ration system, on the other hand, it is easily verified that convertible rations are weakly Pareto superior to nonconvertible rations.

D. *Commodity Taxes and Administrative Costs*

An important generalization of the results presented earlier is that they remain unchanged if there is a tax (or subsidy) on the deficit good, provided the same tax applies under all allocation systems. To see this, let s denote the tax per unit of the deficit good. That is: (*i*) under a government-managed system, the price of the deficit good at government shops is $p + s$; (*ii*) under nonintervention, s is the difference between the market price of the deficit good and the price which firms owning this good receive; and (*iii*) the resulting budget surplus (or deficit) to the government, in each case, is sX per capita. Then, it can be verified that my comparisons among alternative systems are unaffected, regardless of what s is. This is because s cancels out when an individual's utility under alternative systems is compared.

My results are also unaffected by administrative costs, if these costs are not significantly different under alternative systems (i.e., the sum of storage, personnel, and other transaction costs accruing to the government as well as private intermediaries depends primarily on the total quantity of the deficit good), and if these costs are passed on to consumers through the price of the deficit good. This is simply because the effect of administrative cost, in this case, is analogous to that of a commodity tax.

Additional generalizations of the following kind are, therefore, straightforward. Suppose we find that $V^{hI} > V^{hJ}$ when systems I and J are hypothetically assumed to have the same administrative cost, then the same conclusion holds if in fact the system J has a higher administrative cost than that of I. As a specific example, my result that convertible rations are better for the poor than nonconvertible rations holds not only when these two systems entail the same administrative cost, but also when the latter system entails a larger administrative cost (for instance, if the cost of enforcing nonconvertibility exceeds the cost of transacting secondary trades).[12]

IV. Concluding Remarks

Allocation systems such as rationing and queues are extensively employed in many *LDC*s and centrally planned economies. In this paper, I have compared the outcomes of such systems with one another, and with that of unhindered market. My analysis has concentrated on *positive* comparisons: I have attempted to ascertain, for each pair of allocation systems, whether a specific group of individuals (particularly the poor and the rich) is better off under one system or another. The results and insights obtained from these comparisons are valid, as well as informative for policy debates on these issues, regardless of the social criterion or political pressures (resulting, for instance, in an unwillingness to allow the market price to increase) based on which a government might want to choose an allocation system.

I recognize that there is a great diversity in the structures and the economic outcomes of the allocation systems that are employed in different contexts.[13] In this paper, I have used relatively simple models to depict alternative allocation systems and have focused on the comparisons of their outcomes within a narrow but important class of circumstances when the supply of a good is limited.[14] Within this class, however, most of my results are robust not only to parameters such as the cost and the quantity of the deficit good available in the economy, and the nature of heterogeneity in individuals' tastes, but also to certain types of commodity taxes and administrative costs. Moreover, my comparisons among alternative government-managed systems are relevant even when the quantity of the deficit good to be distributed among individuals is a policy choice, rather than a datum for the economy.

[12] Note that this paper does not take a position on whether the total administrative cost under a particular system is larger or smaller than that in another system. This is because the empirical or conceptual basis for such a generalized assertion appears to be inadequate at present. Attention to administrative costs is nevertheless a step in the right direction because these costs are important in practice but, as I indicated earlier, they have been ignored in much of the literature.

[13] See Janos Kornai (1980) for a description of some of the effects of nonprice controls in centrally planned economies; this work, however, does not emphasize a comparison of the outcomes of alternative controls.

[14] Supply responses, on the other hand, have critical implications (for prices as well as individuals' earnings) in many situations. See, for instance, my paper with T. N. Srinivasan (1986) for an analysis of the role of supply responses in determining the distributional consequences of partial food rationing in *LDC* cities.

REFERENCES

Bucovetsky, Sam, "On the Use of Distributional Waits," *Canadian Journal of Economics*, November 1984, *17*, 699–717.

Guesnerie, Roger and Roberts, Kevin, "Effective Policy Tools and Quantity Controls," *Econometrica*, January 1984, *52*, 59–86.

Kornai, J., *Economics of Shortage*, Amsterdam: North-Holland, 1980.

Neary, J. Peter and Roberts, Kevin, "The Theory of Household Behaviour Under Rationing," *European Economic Review*, March 1980, *13*, 25–42.

Sah, Raaj Kumar, "Queues, Rations, and Market: Comparisons of Outcomes for the Poor and the Rich," Economic Growth Center Discussion Paper 504, Yale University, 1986.

______ **and Srinivasan, T. N.,** "Distributional Consequences of Rural Food Levy and Subsidized Urban Rations," Economic Growth Center Discussion Paper 505, Yale University, 1986.

______ **and Stiglitz, Joseph E.,** "The Taxation and Pricing of Agricultural and Industrial Goods in Developing Economies," in D.O.G. Newbery and Nicholas H. Stern, eds., *Modern Tax Theory for Developing*

Countries, Oxford: Oxford University Press, forthcoming.

Tobin, James, "A Survey of the Theory of Rationing," *Econometrica*, October 1952, *20*, 521–53.

Weitzman, Martin L., "Is the Price System or Rationing More Effective in Getting a Commodity to Those Who Need it Most," *Bell Journal of Economics*, Autumn 1977, *8*, 517–24.

Irrelevance of Open Market Operations in Some Economies with Government Currency Being Dominated in Rate of Return

By THOMAS J. SARGENT AND BRUCE D. SMITH*

This paper describes an environment in which government-issued currency is dominated in rate of return and in which there obtains a Modigliani-Miller theorem for government open market operations. Earlier Modigliani-Miller theorems for government finance have been stated for environments in which government-issued currency is not dominated in rate of return in equilibrium. Since government-issued currency is widely observed to be dominated in return, it is useful to study how Modigliani-Miller theorems hinge on absence of rate of return dominance.

Modigliani-Miller theorems for government finance describe environments in which there is a nonsingular equivalence class of government financial policies that support the same equilibrium allocation of goods. Papers by Neil Wallace (1981), Dan Pelèd (1985), and Christophe Chamley and Heraklis Polemarchakis (1984) have explored different dimensions of such an equivalence class of government policies. Wallace characterized a class of government open market exchanges of capital for fiat currency that left the equilibrium sequences for the price level and for real allocations unaffected. Peled characterized a class of government issues of indexed bonds offset by contractions of nominal government debt that left unaltered both real allocations and the price level process. Chamley and Polemarchakis described a government strategy of purchasing capital, financed by alterations in the stock of government issued currency, that left the real allocation unaltered, while systematically altering the price level process so as to pay out to private agents the altered returns on the government's portfolio. The equivalence classes of government financial policies discovered by Wallace, Peled, and Chamley-Polemarchakis can each be thought of as movements in a particular direction within a broader equivalence class of government policies (see Sargent, 1986).

All of these results have been obtained in contexts in which the equilibria being studied are ones in which government-issued fiat currency is not dominated in rate of return (notice the role played by Wallace's equation (4) in his construction). Absence of rate of return dominance has often been ascribed a key role in delivering irrelevance theorems. For example, the failure of currency to be dominated in rate of return was tentatively advanced as a criterion for testing the relevance of government open market operations (Sargent-Wallace, 1983). Sargent and Wallace suggested that the absence of unexploited arbitrage opportunities to the central bank would be a symptom of the irrelevance of central bank open market exchanges, and that the presence of unexploited arbitrage opportunities indicated the necessary relevance of such exchanges.

The presumed link between irrelevance propositions and environments in which rate of return dominance is absent seems to limit the applicability of irrelevance theorems in interpreting observations, given the widely observed inferiority of yield on government issued currency. This point can be used to criticize some of the interpretations advanced in our earlier empirical work (Smith,

*University of Minnesota and Hoover Institution, Stanford University, Stanford, CA 94305–6010, and Carnegie-Mellon University and University of California, Santa Barbara, CA 93106, respectively. We received exceptionally helpful comments from two referees, for which we give thanks. Sargent's research was support by NSF grant SES-8508935 to the University of Minnesota.

1985a, b, and Sargent, 1983), in which we appealed to irrelevance theorems as grounds for believing that increases in government currency that were backed by "real bills" would have no price level effects.

The purpose of this paper is to show that such irrelevance theorems do not necessarily require an environment in which rate of return dominance is absent, nor do they require an environment in which all agents have access to a complete set of state contingent claims markets. We study an environment which combines Wallace's 1981 model with aspects of the Sargent-Wallace 1982 model, in which a proper subset of agents (the "poor") is precluded from private credit markets (as well as markets in state contingent claims) via legal restrictions that are intended to isolate the "money market" from the "credit market." The remaining agents are free to trade in any market they choose. As in Sargent-Wallace (1982), this is an environment in which currency can be dominated in rate of return. For this environment, we show that there obtains an irrelevance theorem with marked similarities to Wallace's. The principal additions to Wallace's hypotheses that are needed to obtain our theorem are that "holding fiscal policy constant" requires making compensating changes in taxes and transfers in a way that respects the restrictions precluding a subset of agents from trading private credit; and that the distribution of income across young agents must be adjusted to insure sufficient demand for currency.

This exploration of the role of rate of return dominance in delivering irrelevance propositions provides ingredients for a sequel (our other paper, 1986) that extends the domain of Modigliani-Miller theorems to government open market operations in foreign currencies. By extending arguments of Wallace, it is possible to obtain irrelevance theorems for exchanges in foreign currencies in environments characterized by Kareken-Wallace exchange rate indeterminacy (John Kareken and Wallace, 1981). (See Rodolpho Manuelli and Sargent, 1986). The Kareken-Wallace environment embodies an absence of rate of return dominance of one government's currency vis-à-vis another's. In our other paper, we study whether and how such irrelevance results can be extended to environments in which exchange rates are determinate and, coincidentally, in which rate of return dominance prevails.

This paper also supplies a reinterpretation of the lump sum taxes and transfers that must be imposed to hold fiscal policy constant. If the menu of assets in which the government trades is broadened sufficiently, then there exists a set of coordinated government open market asset exchanges which hold fiscal policy constant without altering taxes and transfers. This interpretation of holding fiscal policy constant is useful to have because, particularly in the context of Modigliani-Miller theorems for foreign exchange transactions, it is sometimes implausible to imagine that a government has the power literally to impose the offsetting lump sum taxes and transfers on foreign residents that the hypotheses of the theorems seem to require. We show that so long as the government has the ability to issue a sufficiently rich set of state contingent liabilities, there exists a set of asset exchanges which by themselves suffice to hold fiscal policy constant.

I. The Economy

A. *Physical Environment and Markets*

The economy consists of overlapping generations of two-period-lived agents. At each date $t \geq 1$, there is born a set $H(t)$ of agents who are young at t and old at $t+1$. At $t=1$, there also exists a set $H(0)$ of agents who are old. We let $h \in H(t)$ index an individual of generation t. An agent h born at time t is endowed (after taxes and transfers) with $w_t^h(t)$ of a single date t consumption good when young, and $w_{ti}^h(t+1)$ of a single date $t+1$ consumption good when a random variable $x(t+1)=x_i$.[1] The random variable $x(t+1)$ is realized at the beginning of period $t+1$, before time $t+1$ decisions are made,

[1] For notational simplicity we focus on a single-good model. There would be no difficulty in extending the results to a multiple-good context.

but after time t decisions are made. The variable $x(t+1)$ governs the rate of return on storage of the single consumption good between dates t and $t+1$. When an agent h stores $k^h(t)>0$ units of the time t consumption good, "nature" returns $x(t+1)k^h(t)$ units of the time $t+1$ good. We assume that for $t\geq 1$, $x(t+1)$ is nonnegative, is independently and identically distributed over time, and has the discrete probability distribution

$$\text{Prob}\{x(t+1)=x_i\}=f_i, \qquad i=1,\dots,I;$$
$$\sum_{i=1}^{I} f_i=1.$$

Agent h of generation t consumes $c_t^h(t)$ when young, and an x-contingent amount $c_{ti}^h(t+1)$ when old if $x(t+1)=x_i$. Agent $h\in H(t)$; $t\geq 1$, has the objective function

$$Eu^h\left[c_t^h(t),c_{ti}^h(t+1)\right]$$
$$=\sum_{i=1}^{I} f_i u^h\left[c_t^h(t),c_{ti}^h(t+1)\right],$$

where $u^h[c_t^h(t),c_{ti}^h(t+1)]$ is strictly increasing, strictly concave, and twice differentiable.

An exogenous aggregate endowment of the time t good $Y(t)$ is available to the economy $\forall t\geq 1$. It is assumed that $Y(t)$ is nonstochastic. If in the aggregate $K(t)\geq 0$ is stored at t, then $Y(t+1)+x(t+1)K(t)$ of time $t+1$ goods become available. We let $K^p(t)=\Sigma k^h(t)\geq 0$ denote aggregate private storage, where $k^h(t)$ is the amount stored by agent h. We let $K^g(t)\geq 0$ denote government storage. Total storage $K(t)$ is given by $K(t)=K^p(t)+K^g(t)$.

At time 1, there exists a set $H(0)$ of old people who are endowed in the aggregate with $M(0)$ units of an unbacked and inconvertible fiat currency. At $t=1$, an old agent h is endowed after taxes with quantity $w_0^h(1)$ of time 1 consumption good. We assume that $K^g(0)$ is nonnegative and given, and for simplicity that $K^p(0)=0$.

Three kinds of assets are traded in our economy: fiat currency, "storage," and one-period state-contingent claims. We study an economy in which a class of agents is not permitted to trade some of these assets. This prohibition on some trades for some agents separates the "money market" from "credit markets," and creates the possibility that currency is valued even while being dominated in rate of return. We partition $H(t)$, $t\geq 1$, into two disjoint, exhaustive, and nonempty subsets, denoted $H_R(t)$ and $H_P(t)$. Thus $H(t)=H_R(t)\cup H_p(t)$, $H_R(t)\cap H_p(t)=\varnothing$, $H_R(t)\neq\varnothing$, $H_p(t)\neq\varnothing$ for all $t\geq 1$. Agents with $h\in H_R(t)$ are excluded from *no* markets at t, while agents with $h\in H_p(t)$ may only engage in trades of fiat currency for goods.[2,3]

Let $m^h(t)$ denote the nominal amount of fiat currency held by agent h of generation t. We let $p_j(t)$ be the inverse of the price level when $x(t)=x_j$ at time t. We sometimes shall denote the inverse price level at t simply as $p(t)$, with it being understood to depend on j. Thus, the real value of currency stored between t and $t+1$ by agent h is $m^h(t)p(t)$. We let $k^h(t)$ denote the quantity of time t good stored by agent $h\in H(t)$. We let $d_i^h(t)$ denote the number of state-contingent claims for delivery of one unit of time $t+1$ good in state $x(t+1)=x_i$ held by agent $h\in H(t)$. We let $s_i(t)$ be the date t price of a claim to delivery of one unit of the good in state $x(t+1)=x_i$ at $t+1$, denominated in units of the time t good. We assume that the young generation at t is born after the realization of the state $x(t)$, and does not trade claims to state contingent delivery of the good during period t.

All agents in the economy behave competitively and have rational expectations. The

[2] See Sargent-Wallace (1982) for a motivation of this setup in the context of a nonstochastic overlapping generations model with borrowers and lenders. Sargent-Wallace describe a legal restriction on the minimal size of privately issued securities, together with assumptions about the distribution of income, that are sufficient to support an assignment of agents to the classes $H_R(t)$ and $H_p(t)$. Sargent and Wallace's legal restriction was intended to be an abstract representation of the quantity theory restrictions embodied in Peel's Bank Act of 1844.

[3] Agents with $h\in H_p(t)$ are also permitted to hold negative quantities of currency, so that borrowing and lending among members of $H_p(t)$ are permitted at t.

content of our market exclusion restriction is that for all $h \in H_p(t)$, $k^h(t) = d_i^h(t) = 0$ for all $i = 1,\dots,I$ and all $t \geq 1$, and that $m^h(t) \geq 0 \forall h \in H_R(t)$. These restrictions preclude members of $H_p(t)$ from trading in markets for storage or state-contingent claims, and prevent members of $H_R(t)$ from arbitraging between money and credit markets.

B. *Behavior of Private Agents*

"Poor Agents": We first describe the behavior of young agents for whom $h \in H_p(t)$. These agents are allowed to accumulate only fiat currency, and can acquire no other claims on period $t+1$ consumption. For all $h \in H_p(t)$, agent h chooses $(c_t^h(t)$, $c_{ti}^h(t+1)$, $m^h(t))$ to maximize

$$\sum_{i=1}^{I} f_i u^h\left[c_t^h(t), c_{ti}^h(t+1)\right]$$

subject to

$$c_t^h(t) + p(t)m^h(t) \leq w_t^h(t)$$

$$c_{ti}^h(t+1) \leq w_{ti}^h(t+1) + p_i(t+1)m^h(t),$$

$$i = 1,\dots,I,$$

taking $p(t)$, $p_1(t+1)$, $p_2(t+1),\dots, p_I(t+1)$ as given.[4] Solving the second constraint for $m^h(t)$ and substituting into the first constraint gives the following collection of I intertemporal budget constraints, one for each second-period state x_i, that impinge on the poor:

$$c_t^h(t) + \frac{p(t)}{p_i(t+1)} c_{ti}^h(t+1)$$

$$\leq w_t^h(t) + \frac{p(t)}{p_i(t+1)} w_{ti}^h(t+1);$$

$$i = 1,\dots,I.$$

The maximizing choice of $c_t^h(t)$ is described by a demand function

$$(1) \quad c_t^h(t) = g^h\left[p(t), \underline{p}(t+1), w_t^h(t), \underline{w}_t^h(t+1)\right];\ h \in H_p(t),$$

where $\underline{p}(t+1) = [p_1(t+1),\dots, p_I(t+1)]$, and $\underline{w}_t^h(t+1) = [w_{t1}^h(t+1),\dots, w_{tI}^h(t+1)]$.

The choices of $m^h(t)$ and $c_{ti}^h(t+1)$ are given by substituting (1) into the budget constraints

$$(2) \quad p(t)m^h(t) = w_t^h(t) - g^h\left[p(t), \underline{p}(t+1), w_t^h(t), \underline{w}_t^h(t+1)\right]$$

$$(3) \quad c_{ti}^h(t+1) = \frac{p_i(t+1)}{p(t)}\Big[w_t^h(t) - g^h\Big(p(t), \underline{p}(t+1), w_t^h(t), \underline{w}_t^h(t+1)\Big)\Big] + w_{ti}^h(t+1),\ i = 1,\dots,I.$$

"Rich" Agents: We now turn to the rich agents, those with $h \in H_R(t)$. These agents choose $c_t^h(t)$, $c_{ti}^h(t+1)$, $k^h(t)$, $m^h(t)$, and $\underline{d}^h(t) = [d_1^h(t), d_2^h(t),\dots, d_I^h(t)]$ to maximize

$$\sum_{i=1}^{I} f_i u^h\left[c_t^h(t), c_{ti}^h(t+1)\right]$$

subject to

$$c_t^h(t) + k^h(t) + p(t)m^h(t) + \sum_{i=1}^{I} s_i(t) d_i^h(t) \leq w_t^h(t)$$

$$c_{ti}^h(t+1) \leq w_{ti}^h(t+1) + x_i k^h(t) + p_i(t+1)m^h(t) + d_i^h(t),\ i = 1,\dots,I$$

taking $p(t)$, $\underline{p}(t+1)$ and $\underline{s}(t) = [s_1(t),\dots, s_I(t)]$ as given. It is revealing to reformulate this problem as follows. Multiply each side

[4] Again, notice that we have *not* imposed $m^h(t) \geq 0$ for any $h \in H_p(t)$. Hence agents may be "short" in currency, effectively allowing members of $H_p(t)$ to borrow and lend among themselves.

of the second constraint by $s_i(t)$, sum over i, and use the resulting equation to eliminate $\Sigma_i s_i(t) d_i^h(t)$ from the first constraint to obtain

$$
\text{(4)}\quad c_t^h(t) + \sum_{i=1}^{I} s_i(t) c_{ti}^h(t+1) \le k^h(t)\left[\sum_{i=1}^{I} s_i(t) x_i - 1\right] + \left[\sum_{i=1}^{I} s_i(t) p_i(t+1) - p(t)\right] m^h(t) + w_t^h(t) + \sum_{i=1}^{I} s_i(t) w_{ti}^h(t+1).
$$

We henceforth consider the problem of maximizing $\Sigma_i f_i u^h[c_t^h(t), c_{ti}^h(t+1)]$ subject to (4) with $k^h(t) \ge 0, m^h(t) \ge 0$.

Inspection of (4) shows that the absence of arbitrage opportunities for $h \in H_R(t)$ implies that $(\Sigma_{i=1}^I s_i(t) x_i - 1) \le 0$. If $\Sigma_{i=1}^I s_i(t) x_i < 1$, it is evident from (4) that maximizing behavior implies $k^h(t) = 0$. Since we want to discuss Modigliani-Miller theorems that are generated by exchanges of government liabilities for privately held capital, we shall study equilibria in which

$$
\text{(5)}\quad \sum_{i=1}^{I} s_i(t) x_i = 1; \qquad t \ge 1.
$$

We also want to study equilibria in which fiat currency is dominated in rate of return, which will imply that $m^h(t) = 0$ for all $h \in H_R(t)$. A sufficient condition that $m^h(t) = 0$ for all $h \in H_R(t)$ is that

$$
\text{(6)}\quad x_i \ge \frac{p_i(t+1)}{p_j(t)} \qquad \forall i, j, t,
$$

with strict inequality for some i. To show the sufficiency of this condition, suppose to the contrary that $m^h(t) > 0$ for some $h \in H_R(t)$ for some t. Then (4) implies that for $m^h(t) > 0$ and $m^h(t) < \infty$, we require

$$
\text{(7)}\quad \sum_i s_i(t) p_i(t+1) - p_j(t) = 0
$$

for that t, where j is any state consistent with $m^h(t) > 0$ at t. But (5) and (7) imply that

$$
\sum_{i=1}^{I} s_i(t)\left[x_i - \frac{p_i(t+1)}{p_j(t)}\right] = 0.
$$

This contradicts (6), which holds with strict inequality for some i, because $s_i(t) > 0$ for all i, t in equilibrium.[5] Therefore, if (6) holds, $m^h(t) = 0$ for all $h \in H^R(t)$.[6]

When (5) and (6) hold, (4) simplifies to

$$
\text{(8)}\quad c_t^h(t) + \underline{s}(t)\underline{c}_t^h(t+1) \le w_t^h(t) + \underline{s}(t)\underline{w}_t^h(t+1).
$$

Agent $h \in H_R(t)$ maximizes $\Sigma f_i u^h(c_t^h(t), c_{ti}^h(t+1))$ subject to (8). Denote the solution to this problem as

$$
c_t^h(t) = q^h[\underline{s}(t), w_t^h(t), \underline{w}_t^h(t+1)]
$$

$$
c_{ti}^h(t+1) = q_i^h[\underline{s}(t), w_t^h(t), \underline{w}_t^h(t+1)]
$$

$\forall h \in H_R(t)$. Under the assumption that $u^h(\cdot)$ is strictly concave, g^h, q^h, and q_i^h are each functions.

C. *Initial Old*

An initial old agent, $h \in H(0)$, has after-tax endowment $w_0^h(1)$ and holdings of fiat currency $m^h(0)$. Agent $h \in H(0)$ supplies this currency inelastically at $t = 1$. Consumption allocations of the old obey

$$
c_0^h(1) = w_0^h(1) + m^h(0) p(1)
$$

$$
\sum_h c_0^h(1) = \sum_h w_0^h(1) + M(0) p(1)
$$

[5] This depends on the assumption that $f_i > 0$ for all i.

[6] This argument implies that (7) fails to hold. Applying an arbitrage argument to (4) shows that

$$
\sum_{i=1}^{I} s_i(t) p_i(t+1) - p(t) < 0,
$$

in contrast to Wallace's equation (4) (1981). The preceding inequality states that currency is dominated in rate of return.

where $M(0)=\Sigma_h m^h(0)$ is the aggregate endowment of fiat currency of the initial old.

D. *Government*

The government chooses infinite sequences for $t \geq 1$ for the following variables: $M(t)$, the nominal value of the stock of fiat currency at t; $K^g(t)$, the stock of government storage at t; $G_i(t)$, government consumption of the time t good in state i, which leads to no accumulation of subsequent stocks; and $T_t^h(t)$ and $T_{ti}^h(t+1)$, tax receipts from the young and old agents with index h, respectively. Following Wallace (1981), we define

$$w_t^h(t)=y_t^h(t)-T_t^h(t)$$

$$w_{ti}^h(t+1)=y_{ti}^h(t+1)-T_{ti}^h(t+1)$$

so that

$$\sum_{H(t)} y_t^h(t)+\sum_{H(t-1)} y_{t-1,i}^h(t)=Y(t),$$

where $(y_t^h(t), y_{ti}^h(t+1),\ i=1,\dots,I)$ is the pre-tax endowment vector of agent $h \in H(t)$. Recall that the aggregate endowment $Y(t)$ is a nonstochastic sequence. The government chooses its sequences subject to the sequence of budget constraints

$$\begin{aligned} K^g(t)+G_i(t) = & \sum_{H(t-1)} T_{t-1,i}^h(t) \\ & +\sum_{H(t)} T_t^h(t)+K^g(t-1)x_i \\ & +p_i(t)[M(t)-M(t-1)] \end{aligned}$$

for all i,t. The values $M(0)$, $T_i^h(0)$, and $K^g(0)$ are given as initial conditions. We denote total tax collections at t by $T_i(t)=\Sigma_h T_{t-1,i}^h(t)+\Sigma_h T_t^h(t)$.

This concludes our description of the physical environment, the behavior of private agents, and the constraints faced by the government.

II. Equilibrium with Currency Dominated in Rate of Return

We study some properties of an equilibrium in which currency is dominated in rate of return. We use the following definition, which is agnostic about the division of variables between "exogenous" and "endogenous."

Definition: *Given initial conditions,* $M(0)$, $T_i^h(0)$ *and* $K^g(0)$, *a nonrandom nonnegative sequence* $\{Y(t)\}$, *and a stochastic process* $x(t)$ *for* $t \geq 1$, *an equilibrium with currency dominated in rate of return is a collection of stochastic processes for*

$$\begin{gathered} \{K^g(t)\}_{t=1}^{\infty}, \{T(t)\}_{t=1}^{\infty}\{G_i(t)\}_{t=1}^{\infty}, \\ \big[\{w_t^h(t)\}_{t=1}^{\infty}, \{w_{ti}^h(t+1)\}_{t=1}^{\infty}, \\ w_0^h(1),\ h\in H(t)\big], \{M(t)\}_{t=1}^{\infty}, \\ \big[\{c_t^h(t)\}_{t=1}^{\infty}, \{c_{ti}^h(t+1)\}_{t=1}^{\infty}, c_0^h(1), \\ \{k^h(t)\}_{t=1}^{\infty}, \{m^h(t)\}_{t=1}^{\infty}, h\in H(t)\big], \\ \{\underline{p}(t)\}_{t=1}^{\infty}, \textit{and } \{\underline{s}(t)\}_{t=1}^{\infty} \end{gathered}$$

satisfying

$$\sum_h c_0^h(1)=\sum_h w_0^h(1)+M(0)p(1) \tag{9}$$

$$\begin{aligned} c_t^h(t)+\frac{p(t)}{p_i(t+1)}c_{ti}^h(t+1) & \\ \leq w_t^h(t)+\frac{p(t)}{p_i(t+1)}w_{ti}^h(t+1), & \\ \textit{for all } i, \forall h \in H_p(t) & \end{aligned} \tag{10}$$

$$\begin{aligned} c_t^h(t)=g^h\big[p(t),\underline{p}(t+1), & \\ w_t^h(t),\underline{w}_t^h(t+1)\big]\ \forall h\in H_p(t) & \end{aligned} \tag{11}$$

$$m^h(t)p(t)=w_t^h(t)-c_t^h(t),\ h\in H_p(t) \tag{12}$$

$$\sum_{H_p(t)} m^h(t)=M(t), \qquad t\geq 1 \tag{13}$$

$$\begin{aligned} x_i \geq p_i(t+1)/p_j(t) \quad & \textit{for all } i,j,t; \\ & \textit{strict inequality for some } i, \forall j,t \end{aligned} \tag{14}$$

$$(15)\quad c_t^h(t)+\underline{s}(t)\underline{c}_t^h(t+1) \leq w_t^h(t)+\underline{s}(t)\underline{w}_t^h(t+1)\ \forall h \in H_R(t)$$

$$(16)\quad c_t^h(t)=q^h\left[\underline{s}(t), w_t^h(t), \underline{w}_t^h(t+1)\right] \quad \forall h \in H_R(t)$$

$$(17)\quad c_{ti}^h(t+1)=q_i^h\left[\underline{s}(t), w_t^h(t), \underline{w}_t^h(t+1)\right] \quad \forall h \in H_R(t)$$

$$(18)\quad \sum_{i=1}^{I} s_i(t)x_i=1$$

$$(19)\quad \sum_{H_R(t)} c_{ti}^h(t+1) = \sum_{H_R(t)} w_{ti}^h(t+1)+x_i \sum_{H_R(t)} k^h(t)$$

$$(20)\quad K^g(t)+G_i(t)=T_i(t) + K^g(t-1)x_i(t)+p_i(t)[M(t)-M(t-1)]$$

$$(21)\quad T_i(t)=Y(t)-\sum_h w_t^h(t)-\sum_h w_{t-1,i}^h(t)$$

$$(22)\quad T_i(t)=\sum_h T_t^h(t)+\sum_h T_{t-1,i}^h(t).$$

Equation (9) relates the allocation assigned to the initial old people to their endowments, and the value of currency. Equations (10)–(12) describe the constrained optimization for agents $h \in H_P(t)$. Equation (13) is the equilibrium condition in the market for fiat currency. Equation (14) is the condition that assures that fiat currency is dominated in rate of return in equilibrium. Equations (15)–(17) describe the constrained optimization for agents $h \in H_R(t)$ when currency is dominated in rate of return. Equation (18) is the no-arbitrage condition for storage. Equation (19) is the equilibrium condition for second-period consumption for agents $h \in H_R(t)$. (One way to obtain condition (19) is to sum over $h \in H_R(t)$ the second-period budget constraint in state i, and then impose $\Sigma_{H_R(t)}m^h(t)=0$ and $\Sigma_{H_R(t)}d_i^h(t)=0$. The condition $\Sigma_{H_R(t)}d_i^h(t)=0$ is the equilibrium condition in the market for state i-contingent claims.) Equation (20) is the government budget constraint, while (21) and (22) relate taxes to endowments.[7]

III. An Irrelevance Result

Given the preceding definition of equilibrium, Modigliani-Miller theorems for government finance are statements about nonuniqueness of the equilibrium along certain dimensions. These Modigliani-Miller theorems have the following structure. Suppose that an initial equilibrium exists, denoted the $(-)$ equilibrium, and has associated with it a set of allocations, say, $\{\bar{c}_t^h(t), \underline{\bar{c}}_t^h(t+1), \forall h, \forall t \geq 1\}$, $\{\bar{c}_0^h(1), \forall h\}$, $\{\bar{G}_i(t), t \geq 1\}$. Let the government policy sequences associated with this equilibrium be denoted $[\{\bar{M}(t), \bar{K}^g(t)\}, \{\bar{w}_t^h(t), \underline{\bar{w}}_t^h(t+1), \forall h\}, t \geq 1]$, and $\bar{w}_0^h(1)\forall h$. Is this sequence of government policy variables unique in supporting the $(-)$ consumption allocations as an equilibrium? If there is a unique sequence of government policy variables associated with the $(-)$ equilibrium, then any alterations in the government policy sequence will alter the equilibrium consumption allocations. But if the sequence of government policy variables associated with the $(-)$ consumption allocation is not unique, there being an equivalence class of government policies that supports the $(-)$ consumption allocation as an equilibrium, then choices of government policy from among members of this equivalence class are said to be irrelevant.

The structure of such irrelevance results suggests a constructive method for discovering them: fix the consumption allocation associated with an initial equilibrium, impose all of the equilibrium conditions, and solve for the government policy sequences that support the initial consumption allocation. "Irrelevance" results will be attained if a nonsingular equivalence class of government policy sequences is discovered by this procedure.

[7] The definition of equilibrium above contains some conditions that are "redundant." However (9)–(22) conveniently collect conditions which are used in the derivations below.

We shall apply this method to the environment described in Section II. Before proceeding formally, we offer the following heuristic description of what is going on. The government is imagined to conduct an open market operation, simultaneously purchasing capital and issuing currency. We want to compare situations with *identical* paths of government purchases because open market operations are our concern. But an open market exchange alters the government's portfolio and the earnings on it, which means that the change in earnings must somehow be distributed to the public in order to keep the path of government purchases unaltered. Among schemes for distributing these earnings via adjustments in lump sum taxes, there turn out to be many that leave both the equilibrium price system and the allocation unaltered. To support an unaltered equilibrium allocation and price system, lump sum taxes must be adjusted so that 1) the lifetime budget set of each agent, including the government, remains unaltered given an equilibrium price system; 2) the aggregate savings schedule of the poor increases enough to absorb the higher real stock of currency; and 3) aggregate private savings of the rich decrease by the amount of capital purchased by the government. To accomplish 2, the lump sum taxes of the poor must be reduced when they are young, and raised when they are old. To accomplish 3, the lump sum taxes of the rich must be raised when they are young, and reduced when they are old. There exist such taxes that leave the budget set of each agent unaltered at the initial price system. Those taxes and the new government portfolio form an alternative government policy that supports the original equilibrium.

We suppose that there exists an initial equilibrium, denoted the $(-)$ equilibrium. We seek to characterize the class of equilibria, denoted the $(\hat{})$ equilibria, with consumption allocations equal to those of the $(-)$ equilibria. We impose

$$(23)\qquad \hat{c}_t^h(t)=\bar{c}_t^h(t)$$

$$\hat{c}_{ti}^h(t+1)=\bar{c}_{ti}^h(t+1)\qquad \forall t\geq 1, \forall i, \forall h$$

$$(24)\qquad \hat{c}_0^h(1)=\bar{c}_0^h(1)\qquad \forall h.$$

In light of (10)–(12), we can support (23) for $h\in H_p(t)$ by requiring

$$(25)\qquad \frac{\hat{p}(t)}{\hat{p}_i(t+1)}=\frac{\bar{p}(t)}{\bar{p}_i(t+1)},\qquad t\geq 1, i,\ j=1,\dots,I;$$

$$(26)\qquad \hat{w}_t^h(t)+\frac{\hat{p}(t)}{\hat{p}_i(t+1)}\hat{w}_{ti}^h(t+1)=\bar{w}_t^h(t)+\frac{\bar{p}(t)}{\bar{p}_i(t+1)}\bar{w}_{ti}^h(t+1)\qquad \forall i, \forall h\in H_p(t), \forall t\geq 1.$$

In light of (14)–(16), we can support (23) for $h\in H_R(t)$ by requiring

$$(27)\qquad \hat{s}_i(t)=\bar{s}_i(t)\qquad \forall i, \forall t\geq 1;$$

$$(28)\qquad \hat{w}_t^h(t)+\underline{\bar{s}}(t)\underline{\hat{w}}_t^h(t+1)=\bar{w}_t^h(t)+\underline{\bar{s}}(t)\underline{\bar{w}}_t^h(t+1)\qquad \forall h\in H_R(t),\ t\geq 1.$$

In light of (9), we can support (24) by setting[8]

$$(29)\qquad \hat{w}_0^h(1)=\bar{w}_0^h(1)\qquad \forall h;$$

$$(30)\qquad \hat{p}(1)=\bar{p}(1).$$

We suppose that

$$(31)\qquad \hat{Y}(t)=\bar{Y}(t),$$

so that the exogenous endowments are equal across the two equilibria to be compared. By way of fixing allocations, we also impose

$$(32)\qquad \hat{G}_i(t)=\bar{G}_i(t),\ t\geq 1,$$

$$(33)\qquad \hat{K}^g(t)+\hat{K}^p(t)=\bar{K}^g(t)+\bar{K}^p(t),\qquad t\geq 1;$$

$$\bar{K}^g(t)+\bar{K}^p(t)\geq \hat{K}^g(t)\geq 0.$$

[8] Note that by imposing (25) and (30) we are ruling out the kinds of policies studied by Chamley and Polemarchakis.

We now proceed to deduce the restrictions on $\hat{M}(t)$, $\hat{K}^g(t)$, $\hat{w}_t^h(t)$, and $\hat{w}_{ti}^h(t+1)$ that are imposed by (23)–(33) and the definition of equilibrium. Sum the second equation of (23) over $h \in H_R(t)$, then use (31), (33), and (19) to obtain

$$(34) \quad \sum_{H_R(t)} \hat{w}_{ti}^h(t+1) = \sum_{H_R(t)} \bar{w}_{ti}^h(t+1) + x_i\left[\hat{K}^g(t) - \bar{K}^g(t)\right].$$

Equation (34) requires that the government rebate any additional earnings generated by its storage at time $t+1$ to the old members of $H_R(t)$. To deduce the restrictions on $\Sigma_{H_R(t)}\hat{w}_t^h(t)$ implied by (34), multiply both sides of (34) by $s_i(t)$, sum over i, and impose (18) to obtain

$$(35) \quad \sum_{H_R(t)} \underline{\bar{s}}(t)\underline{\hat{w}}_t^h(t+1) = \sum_{H_R(t)} \underline{\bar{s}}(t)\underline{\bar{w}}_t^h(t+1) + \left[\hat{K}^g(t) - \bar{K}^g(t)\right].$$

Summing (28) over $h \in H_R(t)$ and substituting into (35) gives

$$(36) \quad \sum_{H_R(t)} \hat{w}_t^h(t) = \sum_{H_R(t)} \bar{w}_t^h(t) - \left[\hat{K}^g(t) - \bar{K}^g(t)\right], \qquad t \geq 1.$$

Equations (28), (34), and (36) describe the way in which $K^g(t)$ and the endowments for $h \in H_R(t)$ must vary in order to obtain "irrelevance." Similarly, (12), (13), (25), (30), and (23) (for $h \in H_p(t)$) imply that

$$(37) \quad \sum_{H_p(t)} \hat{w}_t^h(t) = \sum_{H_p(t)} \bar{w}_t^h(t) + \bar{p}(t)\left[\hat{M}(t) - \bar{M}(t)\right].$$

Equation (37) describes how, in the aggregate, first-period endowments for members of $H_p(t)$ must be adjusted to assure irrelevance. In particular, the first-period endowments of the poor have to be adjusted so that they will demand an appropriately altered level of real balances at an unchanged price level.[9]

Equations (26), (28), (34), and (37) completely represent the set of restrictions relating the financial policies of the government in the (^) equilibrium to those in the (−) equilibrium. These equations incorporate all of the equilibrium conditions, together with equality of the (^) and (−) allocations. The government budget constraint is satisfied by both the (^) and (−) equilibria, as an implication of Walras' Law.[10]

From the equilibrium policy sequences, many other sequences for $\{w_t^h(t), \underline{w}_t^h(t+1), K^g(t), M(t)\}_{t=1}^{\infty}$ can evidently be constructed. There is thus a nontrivial equivalence class of government policy sequences that support the (−) allocation as an equilibrium. We describe this equivalence class formally in Theorem 1, which we have proved by construction:

THEOREM 1: *Suppose that there exists an initial equilibrium*

$$\{\bar{K}^g(t)\}, \{\bar{T}(t)\}, \{\bar{G}_i(t)\},$$

$$\left[\{\bar{w}_t^h(t)\}, \{\bar{w}_{ti}^h(t+1)\}, \bar{w}_0^h(1), h \in H(t)\right],$$

$$\{\bar{M}(t)\}, \left[\{\bar{c}_t^h(t)\}, \{\bar{c}_{ti}^h(t+1)\}, \bar{c}_0^h(1), \{\bar{k}^h(t)\}, \{\bar{m}^h(t)\}, h \in H(t)\right],$$

$$\{\underline{\bar{p}}(t)\}, \{\underline{\bar{s}}(t)\}.$$

[9] As a counterpart of (34) for the poor, note that (37) and the sum of (26) over h belonging to $H_P(t)$ require that the second-period endowment of the poor satisfy

$$\sum_{H_P(t)} \hat{w}_{ti}^h(t+1) = \sum_{H_P(t)} \bar{w}_{ti}^h(t+1) + \left[\bar{p}_i(t+1)/\bar{p}(t)\right]\left[\bar{p}(t)\bar{M}(t) - \bar{p}(t)\hat{M}(t)\right].$$

[10] At fixed initial equilibrium prices, conditions (26), (28), (34), and (37) describe variations in the endowments of private agents that leave their budget sets unaltered and that satisfy the equilibrium conditions. At those same prices, it follows that the budget set of the last agent (the government) remains unaltered.

Then there exists an equilibrium

$$\{\hat{K}^g(t)\},\{\hat{T}(t)\},\{\overline{G}_i(t)\},$$

$$[\{\hat{w}_t^h(t),\hat{w}_{ti}^h(t+1)\},\hat{w}_0^h(1),h\in H(t)],$$

$$\{\hat{M}(t)\},[\{\bar{c}_t^h(t),\bar{c}_{ti}^h(t+1)\},\bar{c}_0^h(1),$$

$$\{\hat{k}^h(t)\},\{\hat{m}^h(t)\},h\in H(t)],$$

$$\{\underline{\bar{s}}(t)\},\ \textit{and}\ \{\underline{\bar{p}}(t)\},$$

satisfying (23)–(25), (27), (29)–(30), *and* (33), *where* $\{\hat{w}_t^h(t), \hat{w}_{ti}^h(t+1), \hat{K}^g(t), \hat{M}(t)\}$ *is any choice of government policy sequences that satisfies equations* (26), (28), (34), *and* (37).

Conditions (26) and (28) require that each private agent's budget set in the (^) equilibrium equal its budget set in the (−) equilibrium. Thus, (26) and (28) are aspects of holding fiscal policy constant in the sense of holding constant the distribution of wealth across the (^) and (−) equilibria. Equation (34), or equivalently a combination of (36), (28), and (18), requires that the government alter the second-period aggregate endowment of the rich agents so as to distribute to them the altered returns the government makes from holding $\hat{K}^g(t)$ rather than $\overline{K}^g(t)$. In effect, (34) instructs the government to act as a costless intermediary on behalf of the rich. Equation (34) can be interpreted as another aspect of holding fiscal policy constant, requiring that the government distribute its portfolio profits in a way that preserves the distribution of wealth between the rich and poor. It is less transparent that (37) is also part of keeping fiscal policy constant across the (^) and the (−) equilibria. One way to show this is to follow Wallace and define earnings on the government portfolio between $(t-1)$ and t as $E(t)=[x(t)-1]K^g(t-1)-[p(t)-p(t-1)]M(t-1)$, $t\geq 2$. We require that altered earnings on the government portfolio be distributed to private agents:

$$\sum_{H(t-1)}[\hat{w}_{t-1}^h(t-1)-\overline{w}_{t-1}^h(t-1)]$$
$$+\sum_{H(t-1)}[\hat{w}_{t-1,i}^h(t)-\overline{w}_{t-1,i}^h(t)]$$
$$=\hat{E}_i(t)-\overline{E}_i(t).$$

Substituting the definition of $E(t)$ into the above equation and using $\hat{p}(t)=\bar{p}(t)$ gives

$$\sum_{H(t-1)}\hat{w}_{t-1,i}^h(t)-\sum_{H(t-1)}\overline{w}_{t-1,i}^h(t)$$
$$=[x_i(t)-1][\hat{K}^g(t-1)-\overline{K}^g(t-1)]$$
$$-[\bar{p}_i(t)-\bar{p}(t-1)][\hat{M}(t-1)-\overline{M}(t-1)]$$
$$-\sum_{H(t-1)}[\hat{w}_{t-1}^h(t-1)-\overline{w}_{t-1}^h(t-1)],$$
$$t\geq 2,\ i=1,\ldots,I.$$

Use (36) and (34) to eliminate $\hat{K}^g(t-1)-\overline{K}^g(t-1)$ from the above expression; then use the sum of (26) over $H_P(t-1)$ to obtain

$$\sum_{H_P(t-1)}\hat{w}_{t-1}^h(t-1)=\sum_{H_P(t-1)}\overline{w}_{t-1}^h(t-1)$$
$$+\bar{p}(t-1)[\hat{M}(t-1)-\overline{M}(t-1)],\quad t\geq 2.$$

which is (37). Thus, once (34) and (36) have been imposed to protect the wealth of the rich as a whole from being affected by alterations in the government's portfolio, equation (37) is required if there are to be no alterations in the revenues from its portfolio that the government retains.

For the purposes of comparing our setup with Wallace's, we note that (36) and (37) can be added to obtain

$$(38)\quad \bar{p}(t)\hat{M}(t)=\bar{p}(t)\overline{M}(t)$$
$$+[\hat{K}^g(t)-\overline{K}^g(t)]+\sum_{H(t)}[\hat{w}_t^h(t)-\overline{w}_t^h(t)].$$

The government's budget constraint, as described by equations (20)–(22), together with

(32) and (38) imply

$$(39)\quad \sum_{H(t)} \hat{w}^h_{ti}(t+1) = \sum_{H(t)} \bar{w}^h_{ti}(t+1)$$
$$+\left[x_i - \bar{p}_i(t+1)/\bar{p}(t)\right]\left[\hat{K}^g(t) - \bar{K}^g(t)\right]$$
$$-\left[\bar{p}_i(t+1)/\bar{p}(t)\right] \sum_{H(t)} \left[\hat{w}^h_t(t) - \bar{w}^h_t(t)\right].$$

Equation (39) is a version of Wallace's condition (b) (Wallace sets the last term equal to zero, as we are free to do also), and requires the government to distribute the profits from an open market operation. Wallace made no distinction between rich and poor, and his condition (a) imposes (28) across all h belonging to $H(t)$. For Wallace's environment, (28), (38), and (39) characterize an equivalence class of government policies that support a given allocation and given price level process (also see Sargent, 1986). Relative to the equivalence class in Wallace's environment, ours is distinguished by the need to treat rich and poor agents differently in order to preserve the restrictions on market participation that prevent the poor from storing goods and trading state-contingent claims and that prevent the rich from issuing currency. These restrictions on market participation cause us to replace (28) for all agents in Wallace's model with (28) for rich agents and (26) for poor; and to refine Wallace's scheme (39) for distributing earnings on the government portfolio into conditions (34) and (37), which preserve the distribution of wealth between rich and poor.

IV. An Example

We let $H_R(t) = H_R$, $H_p(t) = H_p$ for all t. For all agents $h \in H(t)$ we assume $u^h(c^h_t(t), c^h_{ti}(t+1)) = \ln c^h_t(t) + \beta \ln c^h_{ti}(t+1)$. We let endowments of agents gross of taxes be given by $(y^h_t(t),\ y^h_{ti}(t+1)) = (y^h, 0)$ for all $h \in H(t)$, with $\Sigma_{H(t)} y^h = Y$. We let $x_i \geq 1$, $i = 1, \ldots, I$ hold, with strict inequality for some i such that $f_i > 0$. We display equilibria in which for $h \in H_p$, $w^h_{ti}(t+1) = w^h_{tj}(t+1)$ for all i, j, and in which $p_j(t) = p_i(t)$ for all i, j.

A. *Equilibrium* $(-)$

Consider the policy settings $\bar{K}^g(t) = 0$, $\bar{G}(t) = 0$, $\bar{M}(t) = \bar{M}$, $\bar{w}^h_t(t) = y^h$, $\bar{w}^h_{ti}(t+1) = 0$, $i = 1, \ldots, I$, for all $t \geq 1$. Let $M(0) = \bar{M}$, $K^g(0) = \bar{K}^g(0)$ and $w^h_0(1) = \bar{w}^h_0(1)$ all be given. An equilibrium for the economy is given by

$$\bar{p}(t)\bar{M} = \frac{\beta}{1+\beta}\sum_{H_p} y^h \qquad t \geq 1$$

$$K^p(t) = \bar{K}^p = \frac{\beta}{1+\beta}\sum_{H_R} y^h, \quad t \geq 1$$

$$\bar{s}_i(t) = f_i/x_i.$$

The consumption allocation can be found by substituting these values for $p(t)$ and $s_i(t)$ into agents' demand functions, which are readily derived.

B. *Equilibrium* $(\hat{\ })$

Consider the policy choices $\hat{K}^g(t) = \hat{K}^g > 0$, $t \geq 1$, where $\hat{K}^g \leq \bar{K}^p$;

$$\hat{G}(t) = 0, \quad \hat{M}(t) = \hat{M} = \bar{M} + (1/\bar{p})\hat{K}^g, \qquad t \geq 1,$$

$$\hat{w}^h_0(1) = \bar{w}^h_0(1),\ \hat{M}(0) = \bar{M},\ \hat{K}^g(0) = \bar{K}^g(0),$$

$$\hat{w}^h_t(t) = y^h - [\#H_R]^{-1}\hat{K}^g,\ h \in H_R \quad t \geq 1$$

$$\hat{w}^h_{ti}(t+1) = x_i\hat{K}^g[\#H_R]^{-1},\ h \in H_R \quad t \geq 1$$

$$\hat{w}^h_t(t) = y^h + \left[\#H_p\right]^{-1}\hat{K}^g,\ h \in H_p \quad t \geq 1$$

$$\hat{w}^h_{ti}(t+1) = -\left[\#H_p\right]^{-1}\hat{K}^g,\ h \in H_p \quad t \geq 1.$$

Here "$\#H_R$" denotes the number of rich agents. It can be verified directly that equilibrium values are given by

$$\hat{p}(t) = \bar{p}(t) = \bar{p}, \quad \hat{s}_i(t) = \bar{s}_i(t),$$
$$\hat{K}^p(t) + \hat{K}^g = \bar{K}^p.$$

The consumption allocation is identical with that of equilibrium $(-)$.[11]

V. No Lump Sum Taxation

We have studied an equivalence class of government policies that support the same allocation as an equilibrium even when currency is dominated in rate of return. That class of policies involves government open market exchanges of physical capital for currency that are simultaneously accompanied by alterations in lump sum taxes and transfers. The purpose of this section is to explore whether the ability to actually alter lump sum taxes and transfers is really necessary to obtain irrelevance for open market exchanges. We shall show that the menu of assets traded by the government can be extended in such a way that the alterations in lump sum taxes can be dispensed with. More precisely, with this richer asset structure, there obtains an equivalence class of government policies that support the same equilibrium allocation, with a fixed vector sequence of lump sum taxes. Choices from among this equivalence class of government policies can be regarded as open market strategies in which purchases and sales are made simultaneously in several markets. The government executes these simultaneous trades in order to prevent agents from realizing windfall capital gains as a result of government open market operations. The exchanges can be interpreted as being designed to minimize the disruptions in financial markets associated with open market operations.

We proceed to alter the menu of assets, to show how this alters agents' budget sets, and then to construct an equivalence class of government policies in terms of this new asset structure. We retain the notation of the previous sections, except that now we permit the government to engage in borrowing and lending with private agents. Let $b_t^h(t)$ denote government borrowing (if $b_t^h(t)>0$, lending if $b_t^h(t)<0$) from member h of generation $H(t)$ at t. When the government borrows $b_t^h(t)$ from agent $h\in H(t)$, it repays h $R_i^h(t)b_t^h(t)$ at $t+1$, so that $R_i^h(t)$ is the (gross) rate of interest received by agent h, which is permitted to depend on the realization of $x(t+1)$. The government cannot coerce private lending, so that $R_i^h(t)$ must reflect market interest rates faced by agent $h\in H(t)$. Hence we set

$$(40)\quad R_i^h(t)=x_i;\ \forall h\in H_R(t),\quad \forall i,\forall t\geq 1$$

$$(41)\quad R_i^h(t)=\bar{p}_i(t+1)/\bar{p}(t);\quad \forall h\in H_p(t),\quad \forall i,\forall t\geq 1.$$

With this alteration of the asset structure, the choices faced by individual agents are modified as follows. For agents with $h\in H_R(t)$, the budget constraints are

$$(42)\quad c_t^h(t)+p(t)m^h(t)+k^h(t)+\underline{s}(t)\underline{d}^h(t)+b_t^h(t)\leq w_t^h(t),$$

$$(43)\quad c_{ti}^h(t+1)\leq w_{ti}^h(t+1)+x_ik^h(t)+p_i(t+1)m^h(t)+R_i^h(t)b_t^h(t)+d_i^h(t);\quad i=1,\dots,I.$$

The objective of agent $h\in H_R(t)$ is to maximize $\Sigma f_iu^h[c_t^h(t),c_{ti}^h(t+1)]$ subject to (42) and (43). For agents with $h\in H_p(t)$, the budget constraints become

$$(44)\quad c_t^h(t)+p(t)m^h(t)+b_t^h(t)\leq w_t^h(t)$$

$$(45)\quad c_{ti}^h(t+1)\leq w_{ti}^h(t+1)+p_i(t+1)m^h(t)+R_i^h(t)b_t^h(t);\ i=1,\dots,I.$$

The objective of agent $h\in H_p(t)$ is to maximize $\Sigma_if_iu^h[c_t^h(t),c_{ti}^h(t+1)]$ subject to (44) and (45). The government budget constraint

[11] Notice that in the $(\hat{\ })$ equilibrium the money supply is increased proportionately by $(\hat{M}/\bar{M})$ for all time, and that fiscal policy is "constant" relative to the $(-)$ equilibrium. Nevertheless, the price level sequence is the same in the two equilibria.

becomes

$$(46)\quad K^g(t)+G_i(t)=\sum_{H(t)} T_t^h(t) + \sum_{H(t-1)} T_{t-1,i}^h(t)+x_iK^g(t) + \sum_{H(t)} b_t^h(t) - \sum_{H(t-1)} R_i^h(t-1)b_{t-1}^h(t-1) + p(t)[M(t)-M(t-1)].$$

We have the following

Definition: *Given initial conditions* $M(0)$, $T_i^h(0)$, *and* $K^g(0)$, *a nonrandom nonnegative sequence* $\{Y(t)\}$, *and a stochastic process* $x(t)$ *for* $t\geq 1$, *an equilibrium with currency dominated in rate of return is a collection of stochastic processes for*

$$\{K^g(t)\}_{t=1}^{\infty},\quad \{T(t)\}_{t=1}^{\infty},\quad \{G_i(t)\}_{t=1}^{\infty},$$

$$\left[\{w_t^h(t)\}_{t=1}^{\infty}, \{w_{ti}^h(t+1)\}_{t=1}^{\infty}, \{b_t^h(t)\}_{t=1}^{\infty},\right.$$

$$\left. w_0^h(1),\ h\in H(t)\right], \{M(t)\}_{t=1}^{\infty},$$

$$\left[\{c_t^h(t)\}_{t=1}^{\infty}, \{c_{ti}^h(t+1)\}_{t=1}^{\infty}, c_0^h(1),\right.$$

$$\left.\{k^h(t)\}_{t=1}^{\infty}, \{m^h(t)\}_{t=1}^{\infty},\ h\in H(t)\right],$$

$$\{\underline{p}(t)\}_{t=1}^{\infty},\quad \textit{and}\quad \{\underline{s}(t)\}_{t=1}^{\infty}$$

satisfying (9)–(11), (13)–(18), (46), (21)–(22), *and the following two conditions*:

$$(47)\quad m^h(t)p(t)=w_t^h(t)-c_t^h(t)-b_t^h(t);\qquad h\in H_p(t)$$

$$(48)\quad \sum_{H_R(t)} c_{ti}^h(t+1)=\sum_{H_R(t)} w_{ti}^h(t+1) + x_i\sum_{H_R(t)} k^h(t)+R_i^h(t)\sum_{H_R(t)} b_t^h(t).$$

We focus on a $(-)$ equilibrium, and produce conditions under which there is a $(\hat{\ })$ equilibrium satisfying (23)–(25), (29)–(33), (27), $\hat{w}_t^h(t)=\overline{w}_t^h(t)$, and $\hat{w}_{ti}^h(t+1)=\overline{w}_{ti}^h(t+1)\forall i,\ t\geq 1,\ h\in H(t)$. We can support (23) for $h\in H_p(t)$ by requiring (26). We can support (23) for $h\in H_R(t)$ by requiring (28). It remains only to find conditions under which $\hat{K}^g(t)+\hat{K}^p(t)=\overline{K}^g(t)+\overline{K}^p(t)$, and under which $\bar{p}(t)\Sigma_{H_p(t)}\hat{m}^h(t)=\bar{p}(t)\hat{M}(t)$.

In order to satisfy the first of these conditions, sum the second equation of (23) over $h\in H_p(t)$, and use (31), (33), and (48) to obtain

$$(49)\quad R_i^h(t)\sum_{H_R(t)}\left[\hat{b}_t^h(t)-\bar{b}_t^h(t)\right] = x_i\left[\hat{K}^g(t)-\overline{K}^g(t)\right],\quad i=1,\dots,I,$$

where we have imposed constancy of after-tax endowments. Using (40), we see that (49) is equivalent to

$$(50)\quad \sum_{H_R(t)}\left[\hat{b}_t^h(t)-\bar{b}_t^h(t)\right]=\hat{K}^g(t)-\overline{K}^g(t).$$

Similarly,

$$(51)\quad \bar{p}(t)\hat{m}_t^h(t)=\hat{w}^h(t)-\hat{c}_t^h(t)-\hat{b}_t^h(t);\qquad h\in H_p(t)$$

$$(52)\quad \bar{p}(t)\overline{m}_t^h(t)=\overline{w}^h(t)-\bar{c}_t^h(t)-\bar{b}_t^h(t);\qquad h\in H_p(t).$$

Subtracting (52) from (51), using (23), and using $\hat{w}_t^h(t)=\overline{w}_t^h(t)$ yields

$$(53)\quad \bar{p}(t)\left[\hat{M}(t)-\overline{M}(t)\right] = \sum_{H_p(t)}\left[\bar{b}_t^h(t)-\hat{b}_t^h(t)\right].$$

Thus (50) and (53) describe restrictions on $\{\hat{b}_t^h(t)\}$, $\{\hat{M}(t)\}$, and $\{\hat{K}^g(t)\}$ which deliver irrelevance. It is straightforward to verify that sequences satisfying these conditions also satisfy the government budget constraint $\forall t$, and that versions of equations (38) and (39) hold. It follows that a version of Theorem 1 holds in which (50) replaces condition (34) and (53) replaces condition (37). This version of Theorem 1 is one in which government open market exchanges of physical capital are irrelevant when accompanied by a set of government exchanges in debt markets that leave private agents' budget sets unaltered.[12]

This reinterpretation of our theorem is useful in the context of the sequel to this paper, which studies government open market purchases of a foreign currency. In that context, it is especially useful to understand the sense in which irrelevance hinges on the ability of a government to levy lump sum taxes on foreign residents.

VI. Conclusions

We have studied an environment and a structure of restrictions on trades for which government-issued currency is dominated in rate of return and for which open market operations are irrelevant. As did Wallace's earlier result, irrelevance requires that fiscal policy be held constant in a precise sense. Holding fiscal policy constant in the face of a government asset exchange can be thought of as *defining* an open market operation. The hypotheses of our irrelevance theorem are modified relative to those of Wallace's only because we have modified the opportunity sets of a subset of agents in order to create a captive demand for fiat currency that permits it to be valued even as it is dominated in rate of return.

We do not claim that such irrelevance results are obtained in all contexts in which fiat currency is dominated in rate of return. It all depends on the nature of the legal restriction that induces a captive demand for government currency. For example, if that captive demand is generated by imposing a reserve requirement that treats all agents symmetrically, as in Wallace (Section VI), irrelevance results cannot be obtained for government asset exchanges that look like open market operations. To obtain irrelevance theorems seems to require that the structure which generates a demand for government currency be one that impinges differentially on different classes of agents.

In our model, currency is dominated in rate of return, but the *same* agent never holds both currency and interest-bearing claims to future currency. A feature of the reserve requirement analyzed by Wallace is that in equilibrium it can occur that the same agent does hold both currency and other assets that dominate it in rate of return. What precludes a Modigliani-Miller theorem in Wallace's environment is not that feature of the equilibrium, but the imposition of identical reserve requirements on all agents in the economy. If different reserve requirements are imposed on different agents, there can occur Modigliani-Miller theorems in versions of Wallace's model with reserve requirements. This claim can be substantiated by applying the constructive method of Section III of this paper to a version of Wallace's model that has been modified to embody heterogeneous reserve requirements on different agents.

REFERENCES

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November-December 1974, *82*, 1095–117.

Chamley, Christophe and Polemarchakis, Heraklis, "Assets, General Equilibrium, and the Neutrality of Money," *Review of Economic Studies*, January 1984, *51*, 129–38.

Kareken, John and Wallace, Neil, "On the Indeterminacy of Equilibrium Exchange Rates," *Quarterly Journal of Economics*, May 1981, *96*, 207–22.

Manuelli, Rodolfo and Sargent, Thomas J., *Exercises in Dynamic Macroeconomic Theory*, Cambridge: Harvard University Press, 1986.

Modigliani, Franco and Miller, Merton, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review*, June 1958, *48*, 261–97.

Peled, Dan, "Stochastic Inflation and Government Provision of Indexed Bonds," *Journal of Monetary Economics*, May 1985,

[12] The equivalence of these borrowing/taxation schemes with pure intertemporal tax-transfer schemes is one theme of Robert Barro (1974).

15, 291–308.

Sargent, Thomas J., "The Ends of Four Big Inflations," in R. Hall, ed., *Inflation: Causes and Effects*, NBER, Chicago: University of Chicago Press, 1983.

______, *Dynamic Macroeconomic Theory*, Cambridge: Harvard University Press, 1986.

______, **and Smith, Bruce D.,** "The Irrelevance of Government Foreign Exchange Operations," manuscript, 1986.

______ **and Wallace, Neil,** "The Real Bills Doctrine vs. the Quantity Theory: A Reconsideration," *Journal of Political Economy*, December 1982, *90*, 1212–36.

______ **and** ______, "A Model of Commodity Money," *Journal of Monetary Economics*, July 1983, *12*, 163–87.

Smith, Bruce D., (1985a) "Some Colonial Evidence on Two Theories of Money," *Journal of Political Economy*, December 1985, *93*, 1178–211.

______, (1985b) "American Colonial Monetary Regimes: the Failure of the Quantity Theory and Some Evidence in Favour of an Alternate View," *Canadian Journal of Economics*, August 1985, *18*, 531–65.

Wallace, Neil, "A Modigliani-Miller Theorem for Open Market Operations," *American Economic Review*, June 1981, *71*, 267–74.

Exchange Rates and Prices

By RUDIGER DORNBUSCH*

The adjustment of relative prices to exchange rate movements is explained in an industrial organization approach, using various models. The extent of price adjustment, given labor costs in the respective currencies, is shown to depend on product substitutability, the relative number of domestic and foreign firms, and market structure. Some empirical evidence is offered to support the theory.

The large appreciation of the U.S. dollar over the 1980–85 period and the subsequent depreciation open important areas of research. The fact of a large and persistent real appreciation poses a challenge for equilibrium theorists to uncover the change in fundamentals. For those who explain medium-term macroeconomics in terms of Fischer-Taylor long-term wage contracts, the episode provides a striking example of the differential speeds of adjustment of wages, goods, and assets prices. I adopt this perspective here to explain the determinants of relative price changes of different groups of goods. Specifically, I advance hypotheses about those sectors where an exchange rate change should lead to large relative price changes, and others where the relative price effects should be negligible.

The approach is to draw on models of industrial organization to explain price adjustments in terms of the degree of market concentration, the extent of product homogeneity and substitutability, and the relative market shares of domestic and foreign firms. Models of industrial organization have, of course, been fruitfully applied in trade theory; their application to macro-pricing issues, however, has been surprisingly slow.[1] There is a long-standing questioning of purchasing power parity (*PPP*), especially in the work of Irving Kravis and Robert Lipsey (1978, 1984).[2] But so far there seems to exist little formal analysis of price-setting behavior in this context.[3]

This paper adopts a partial-equilibrium approach in that it assumes throughout a given, exogeneous movement in the nominal exchange rate. The exchange rate movement and the less-than-fully flexible money wage interact to produce a cost shock for some firms in an industry—foreign firms in the home market and home firms abroad—and thus bring about the need for an industrywide adjustment in prices. Although the assumption of exogeneous exchange rate movements and sticky wages is open to criticism, it is a useful working hypothesis for the purpose of investigating relative price issues.

Section I reviews some facts. Section II offers a stylized view of the link between exchange rates and prices, and the third

*Massachusetts Institute of Technology, Cambridge, MA 02139. I am indebted to Olivier Blanchard, Stanley Fischer, Paul Krugman, Michael Rothschild, Julio Rotemberg, Sergio Sanchez, Lawrence Summers, and Jean Tirole. Avinash Dixit provided especially valuable suggestions. An earlier version of this paper was presented at the 1985 NBER Summer Workshop and the NBER Meeting on Business Fluctuations and I acknowledge helpful comments received on those occasions. David Wilcox provided valuable research assistance.

[1] See Avinash Dixit (1984) and Elhanan Helpman and Paul Krugman (1985) for extensive work on and references to trade applications. In the macro context, see Olivier Blanchard (1985), Oliver Hart (1982), and N. Gregory Mankiw (1985).

[2] For a review of the *PPP* literature, see my papers (1985, 1986).

[3] Joshua Aizenman (1985, 1986) and Alberto Giovannini (1985) investigate price-setting behavior in the context of exchange rate movements. Their focus, however, is on short-term issues of transactions costs and uncertainty rather than on the large, persistent movements in the real exchange rate. See, too, the more recent papers by Catherine Mann (1986), Robert Feinberg (1986), and Eugene Flood (1986).

section studies the behavior of equilibrium prices.

I. Some Facts

The large dollar movements are reflected both in absolute and relative prices. Table 1 shows two measures of the change in U.S. relative costs and prices: relative unit labor costs and the relative value-added deflator in manufacturing. In each case, the U.S. series is deflated by the corresponding time-series for the trade-weighted average in dollars of our trading partners. The large magnitude of the change in relative costs and prices arises from the fact that unit labor costs and prices abroad (in national currencies) were rising at a lower rate than in the United States. But at the same time, the dollar, rather than offsetting the divergent trend by a depreciation, further reinforced that divergence by a strong appreciation. The depreciation since mid-1985 has not been sufficient to eliminate the change in competitiveness.

Figure 1 shows absolute prices measured by the U.S. GNP deflator and the deflators for imports. Prior to 1980, import prices increase more rapidly than the deflator and, to a lesser extent, so do export prices (not shown). During this period, the dollar was depreciating. After 1980, however, the dollar appreciation gets underway, and import price increases slow down and ultimately import prices fall in absolute terms. Export prices track the GDP deflator more closely, though the pattern of divergences is similar to that for imports. At this broad level, it is clear then that import prices fell relative to the deflator and relative to export prices.

In the absence of comprehensive price series, Table 2 shows unit values for different export and import groups. The table brings out that the absolute decline in import prices must be primarily attributed to the first three groups, and not to finished manufactures. Oil price increases in 1979 easily explain the divergent pattern of export and import unit values for crude materials. The interesting comparison, therefore, is between the relatively homogeneous commodity groups—food and semimanufactures—and finished manufactures where price setting and product differentiation are likely to be important. For the former group, export and import unit values move roughly in line, while for finished manufactures, exports follow the domestic price trend and imports show a much smaller increase.

TABLE 1—RELATIVE COSTS AND PRICES IN MANUFACTURING
(Cumulative Percentage Change)

	1976–80	1980–85:I	1980–85:IV
Relative Value-Added Deflator	−14.7	49.3	27.0
Relative Unit Labor Cost	−12.6	59.8	22.0

Source: International Finance Statistics, Yearbook 1985 and August 1986.

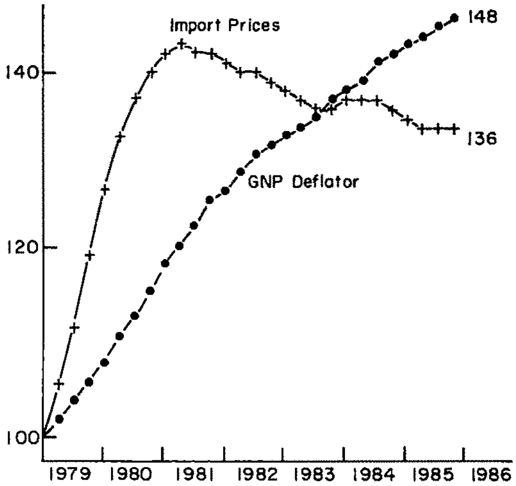


FIGURE 1. THE IMPACT OF DOLLAR APPRECIATION

II. Standard Models

There are two extreme models of price relationships in the open economy literature. One assumes that the "law of one price" holds. Prices of goods are geographically arbitraged and, adjusted for tariffs and transport costs, they are equalized in different locations. Homogeneity, information and perfect competition assure this result.[4] Let

[4] For a review of *PPP*, see my paper (1985).

TABLE 2—UNIT VALUES OF IMPORTS AND EXPORTS (Index 1980:I = 100)

	Foods		Materials		Semi-manufactures		Finished manufactures	
	E	*M*	*E*	*M*	*E*	*M*	*E*	*M*
1979:II	87	82	92	60	71	77	95	91
1980:I	100	100	100	100	100	100	100	100
1985:I	94	87	91	97	86	82	139	106

Note: E = Exports, M = Imports; for definition and source, see the Appendix.

p_i, p_i^*, and e denote the price of good i in the home country and currency, the foreign price, and the home-currency price of foreign exchange. Arbitrage then implies

$$p_i = ep_i^*. \tag{1}$$

In this form, or in the first-difference version of Gustav Cassel, the law of one price is asserted in the *PPP* literature. The law of one price has been applied in the monetary approach to exchange rates in combination with the quantity theory of money and the assumption of full-price flexibility to obtain a theory of the exchange rate. An important implication of complete spatial arbitrage, not only for commodities but for all goods, is the idea that relative national price levels in a common currency are independent of the exchange rate, since exchange rate movements merely reflect, passively, divergent national price trends. That is, of course, an application of the homogeneity postulate which holds when money is fully neutral.

The alternative model might be called "Keynesian." Here it is assumed that each country is fully specialized in the production of "its own" good. Domestic and foreign goods are less than fully homogeneous or substitutable. Wages are fixed in national currencies or at least sticky.

Letting P and P^* be the national GDP deflators, the relative price of domestic and foreign goods or the real exchange rate then, is

$$\lambda = P/eP^*. \tag{2}$$

If the markup of prices over unit labor costs is constant, then for given unit labor costs, prices will also be given. Hence in this model, exchange rate movements change relative prices one-for-one. Exchange rate-induced changes in the relative price affect the world distribution of demand and employment. This approach tends to be used in open economy versions of the IS-LM model in the Meade-Mundell tradition.

Equation (1) would be a useful model of international price relations for materials—say sisal, copper, tea—whereas (2) more nearly describes what happens with manufactures. But the assumption of a constant markup is no longer justified as we shall now see when domestic and foreign firms have strategic interactions in their pricing.

III. Equilibrium Pricing Models

Table 1 gave evidence of large, persistent fluctuations in exchange rate-adjusted relative prices in manufacturing. In this section I explore theoretical models that would explain these price movements as the result of changes in relative unit labor costs.

The basic assumption I make is that firms in any industry have a linear technology, with labor as the only input. Unit labor costs, w and w^*, are given in home and foreign currency, respectively. This assumption about costs is combined with a model of pricing to yield predictions about the behavior of relative prices. The experiment is simply this: the exchange rate change, say a dollar appreciation, lowers foreign unit labor costs in dollars. As a result, the market equilibrium is disturbed in each industry and price and output adjustments must occur.

What these adjustments look like depends on three factors:[5]

Market integration or separation. Is a particular commodity traded in an integrated world market, or are there significant barriers to restrict spatial arbitrage?

Substitution between domestic and foreign variants of a product. The extent of substitution influences price setting and the output effects of cost and price changes.

Market organization. Is the market perfectly competitive, in which case firms are price takers, or is the market imperfectly competitive or oligopolistic, in which case firms are price setters and may interact in strategic ways?

Two models lend themselves in a straightforward fashion to formulating the price response to cost shocks of part of the industry. The Cournot model assumes perfect substitution between alternative suppliers and places more emphasis on the extent of oligopoly. It allows in principle more variation in the markup in response to cost shocks, and thus has the potential for a richer pattern of response to cost shocks. The Dixit-Stiglitz (1977) model by contrast emphasizes imperfect substitution between alternative suppliers and in its predictions looks very much like the Keynesian model discussed above. An alternative to the Dixit-Stiglitz model, again emphasizing product differentiation, is the Salop (1979) model of competition on a circle.

A. *The Cournot Model*

In the Cournot formulation, the analytical focus is on a homogeneous commodity sold in an oligopolistic market. Each seller assumes that other sellers defend their sales volume. I assume that there is an effective spatial separation between the home market and foreign markets, and discuss the pricing in the U.S. market.

For expository purposes, I start with the case of a linear demand function:

$$Q_d = a - bp, \tag{3}$$

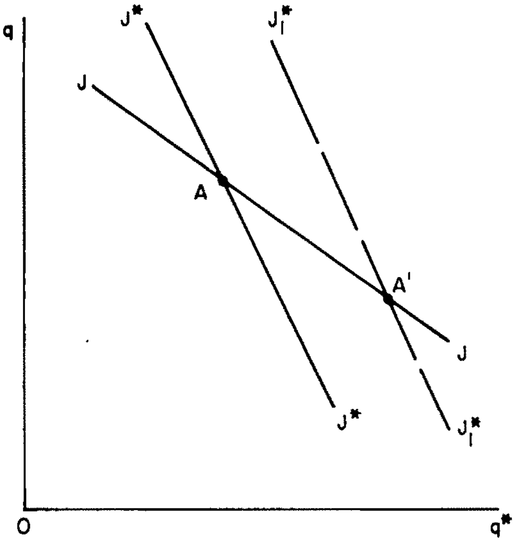


FIGURE 2. THE COURNOT EQUILIBRIUM

where all nonprice determinants are captured in the constant. There are n domestic suppliers and n^* foreign firms with respective sales of q and q^* per firm. Aggregate sales of these firms, Q, have to sum to market demand:

$$Q = nq + n^*q^*. \tag{4}$$

Each firm maximizes profits taking the sales of other firms as given. Profits of the representative domestic and foreign firm in the home market are

$$\begin{aligned} \pi_i &= (p - w) \\ &\quad \times [a - bp - (n-1)q - n^*q^*], \\ \pi_j &= (p/e - w^*) \\ &\quad \times [a - bp - nq - (n^* - 1)q^*]. \end{aligned} \tag{5}$$

Maximization gives rise to the reaction functions shown in Figure 2. The home country's reaction function is JJ, while J^*J^* represents the foreign country. They yield the Cournot-Nash equilibrium shown at point A which gives the equilibrium quantity allocation between representative domestic and foreign firms. The common equilibrium

[5] A fourth item of relevance is the functional form of the demand curve. See Jesus Seade (1983) and Dixit (1986).

price in the industry is given by

$$(6)\quad p=(nw+n^{*}ew^{*})/N+a/bN;$$

$$N\equiv n+n^{*}+1.$$

A dollar appreciation shifts the $J^{*}J^{*}$ schedule out and to the right, thus leading to increased foreign sales and reduced domestic sales. At the initial level of sales for every other supplier, the individual foreign firm faces a given marginal revenue schedule in dollars but experiences a reduction in its dollar marginal cost, and hence wishes to increase output. In the new equilibrium at point A', foreign firms increase their output while home firms contract. The industry price declines, as seen from (6).

We are now interested in the extent to which exchange rate movements (or movements in relative unit labor costs) affect the equilibrium price. The elasticity of the equilibrium price with respect to the exchange rate is

$$(7)\qquad \varphi=(n^{*}/N)(ew^{*}/p).$$

The elasticity formula has two determinants: the relative number of foreign firms (or the relative number of firms with wages not fixed in dollars), and the ratio of marginal cost to price of foreign suppliers. Since both terms are fractions, it is immediately clear from (7) that a dollar appreciation will lower price less than proportionally. The decline in the dollar price is larger the more competitive the industry (i.e., the smaller the markup of price over marginal cost) and the larger the share of imports in total sales. This latter term is represented by n^{*}/N on the assumption of symmetry and initially equal wages between countries.

Equation (7) is interesting because it stretches all the way from the "small country" case to the case where exchange rates have virtually no impact on home pricing. The small country case, in the trade literature, is the case where a country is a price taker in world markets. In that case, a currency depreciation will raise prices in the same proportion. This is, of course, the limiting case for perfect competition *and* a large number of foreign firms relative to the number of home firms.

The other extreme case where exchange rate has no influence on home prices results when there are few firms in the industry, most of which are domestic. In that case, foreign firms absorb the dollar appreciation primarily in the form of extra profits rather than increased sales.

The Cournot model thus potentially explains both unchanging prices and steep price declines. The market structure—import share and concentration—are the key parameters that explain the outcome.

Consider next U.S. export firms competing in a foreign market. A dollar appreciation will lower their marginal revenue in dollars. With unchanged marginal dollar cost, these firms will contract. In terms of Figure 2 applied to the foreign market, our schedule JJ shifts down and to the left. The common foreign currency price rises, but in dollars it declines, though less than proportionately to the appreciation. Using the same model for the foreign market, we find that the elasticity of foreign price with respect to the exchange rate is

$$(8)\qquad \varphi^{*}=-(n'/N^{*})(w/ep^{*}),$$

where n' is the number of domestic firms in the foreign market and N^{*} the total number of firms. With φ^{*} a negative fraction, the dollar price of exports, $p^{*}e$, has an elasticity $1+\varphi$, and hence must decline in response to a dollar appreciation.

Remembering that the markets are separated, let us look at the price of U.S. exports relative to the price of imports p/ep^{*}. In case of a dollar appreciation, dollar export prices rise relative to import prices if the following condition holds:

$$(9)\qquad \varphi>1+\varphi^{*}.$$

In principle, the condition can go either way. In the small country case, export and import prices in dollars fall in the same proportion as the currency appreciates ($\varphi=1$, $\varphi^{*}=-1$), so that the relative price p/ep^{*} remains constant. In general, the outcome depends on the relative oligopolistic structure of the two markets. Export prices will rise relative to import prices in the appreciation case, if at-home import competition is pervasive and foreign markets are strongly affected by U.S. suppliers as well as highly competitive.

Jesus Seade (1983) and Avinash Dixit (1986) have shown that the price effects of disturbances in oligopoly models are highly sensitive to the functional form. It is therefore interesting to ask what happens with a more general functional form. For example, with a constant elasticity demand function $D = Ap\exp(-\beta)$, the elasticity of equilibrium price with respect to the exchange rate becomes $\varphi = (n^*/N)(ew^*/W)$, where $N = n + n^*$ and $W = (n/N)w + (n^*/N)ew^*$. The exchange rate impact thus depends no longer on cost-price markup and, when costs are initially equal between countries, is only a function of the relative number of firms.

In general the elasticity of price with respect to the exchange rate is[6]

$$(9') \quad \varphi = [n^*/(N-\theta)](ew^*/p) > 0,$$

where θ is the elasticity of the slope of the inverse demand curve which, for stability, is less than N. In the linear case, the formula reduces to (9) since $\theta = 0$. But, since θ is positive (or zero), there is no upper bound on φ. From public finance applications, it is known that tax disturbances may lead to magnified price adjustment under oligopoly. I find the same possibility here for the exchange rate effect on prices.

B. *The Dixit-Stiglitz Model*

The representative consumer in this model maximizes a utility function V with consumption of two commodities z and x as arguments:

$$(10) \quad V = U(z,x); \quad x = \left(\sum x_i^a\right)^{1/a}, \quad 0 < a < 1.$$

I focus on commodity x which is an index of consumption of different brands of the same good. It is assumed that there are n domestic firms supplying some variant each, and n^* foreign firms doing the same.

Maximization yields the demand for each individual brand, as well as the utility-based price index for commodity x:

$$(11) \quad x_i = x(P/p_i)^c; \quad c = 1/(1-a),$$

$$(12) \quad P = \left[\left(\sum p_i^h + \sum p_j^h\right)\right]^{1/h}; \quad h = -a/(1-a).$$

In equation (12) p_i denotes the price of a brand produced in the home country, p_j is the price of an imported brand, and P is the industry price.

We are now interested in the response of prices to cost shocks. The individual imperfectly competitive firm faces a demand curve as in (11) with the relative price of its product p_i/P as the determinant. The firm assumes it is sufficiently small so that its own price changes leave the industry price, P, unchanged. The representative firm's profits are

$$(13) \quad \pi_i = (p_i - w)x_i.$$

Maximization yields the familiar constant markup pricing equation:

$$(14) \quad p_i = \alpha w; \quad \alpha = 1/(1-1/c),$$

where α depends inversely on the elasticity of substitution among variants. Since the industry structure is symmetric, each domestic firm will follow the same pricing rule with an equal markup.

Now assume again markets are separated and we can thus meaningfully discuss the price set by a foreign firm for our market. Foreign firms in the home market face the

[6] Let $p = F(Q)$ be the inverse market demand curve. The elasticity of the slope of the inverse demand curve then is $\theta = -QF''(Q)/F'(Q)$. The derivation of (9′) is as follows: The individual firm maximizes profits $\pi = F(Q)q - wq$ and $\pi = (F(Q)/e)q^* - w^*q^*$. The first-order conditions are $F(Q) + qF'(Q) = w$ and $F(Q) + q^*F'(Q) = ew^*$. Multiplying the first-order conditions respectively by n and n^* and adding them yields: $(n+n^*)F(Q) + QF'(Q) = nw + n^*ew^*$. Differentiating this expression, using $dp = F'(Q)dQ$, yields the expression for θ in (9′).

same form of demand curve as home firms and hence they also follow the same pricing rule, with the same markup, but with foreign wages in dollars, ew^*, as the base of their pricing:

$$p_j = \alpha e w^*. \tag{15}$$

From (14) and (15), we have two strong predictions: First the relative price of domestic and foreign variants in the home market depends just on relative unit labor costs in a common currency:

$$p_i/p_j = w/ew^*. \tag{16}$$

Second, it is readily shown that the relative price of a domestic variant in terms of the industry price index (p_i/P) is just a function of the relative wage, w/ew^*. The elasticity of the relative price will be

$$n^*z/(n+n^*z); \quad z=(w/ew^*)^{1/h}. \tag{17}$$

If wages are initially equal between countries, the effect of an exchange rate change on the industry price and on the relative price depends merely on the fraction of firms that has wages fixed in foreign currency, and hence experiences a reduction of their costs in dollars when the dollar appreciates.

Given the wages in home and foreign currency, the Dixit-Stiglitz model provides strong predictions about the impact of dollar appreciation:

The prices of imported variants fall in proportion to the decline of dollar unit labor costs of foreign firms and the prices of domestic variants would remain unchanged.

Exporting firms at home, although they have to compete in foreign markets, still follow their markup pricing on dollar wages. Accordingly, a change in the dollar does not affect their dollar export price. Of course, it does affect their sales and profits. A dollar appreciation will raise their foreign currency price in the same proportion and hence raise their relative price in the foreign market.

The strong prediction of the model is to look for a sharp fall in import prices relative to domestic prices and to see export prices stay constant relative to domestic prices of the same variant. This is, of course, the exact specification of the fixed-price Keynesian model which is derived here as an implication of given labor costs and an invariant markup.

C. *An Extended Dixit-Stiglitz Model*

The Dixit-Stiglitz model assumes Chamberlinian imperfect competition and hence each supplier assumes that he does not affect industry price. Strategic interaction with other firms is therefore excluded. But the same structure of differentiated products can easily be adapted to introduce strategic interaction by way of a conjectural variation. Assume, contrary to the preceding section, that the individual firm is sufficiently large to affect industry price. Assume, too, that firms respond to changes in the industry price, and let the conjectural variation be the parameter σ, a fraction between zero and one. Thus a one-percentage-point rise in the industry price is assumed to cause each firm to raise their price by σ percent. Assuming a given conjectural variation rather than deriving it from a dynamic game-theoretic framework is obviously a shortcut. Nor is there any concern here with consistent conjectural variations.

With this adaptation, the demand curve facing the individual firm's price policy no longer is a constant markup over unit labor costs but rather becomes

$$p_i = \alpha' w; \quad \alpha' \equiv 1/[1-1/c(1-\varepsilon)], \tag{18}$$

where the term $\varepsilon \equiv (dP/P)/(dP_i/P_i)$ captures the strategic interaction between firms as perceived by the individual price-setting firm. The term is a function of relative prices and the conjectural variation:[7]

[7] The derivation is as follows: maximization of profits for a domestic firm k in (13) yields $1+(p_k-w)c(\varepsilon-1)/p_k=0$ which yields the markup equation in (18). The elasticity ε of the aggregate price level with respect to a variation in an individual domestic price p_k is derived as follows: from (12) we have $p^h = p_k^h + \Sigma p_{i\neq k}^h + \Sigma p_j^h$. Now assuming that $\sigma = d\ln p_i/d\ln P$ for all firms i and j other than k we obtain the elasticity of aggregate price with respect to a variation in p_k.

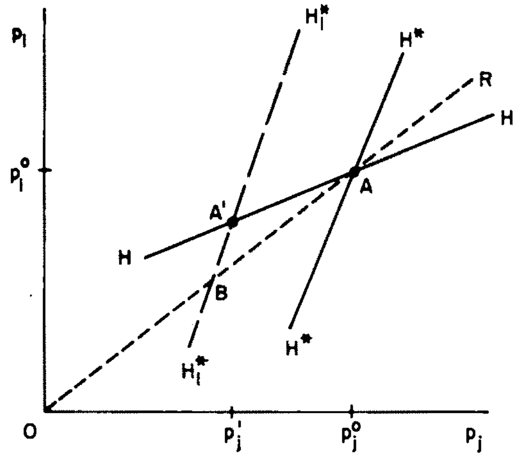


FIGURE 3. THE EXTENDED DIXIT-STIGLITZ MODEL

$$(19)\quad 0<\varepsilon(\sigma, p_i/p_j)\pi \equiv \left[\sigma+(1-\sigma)\left\{n+n^*(p_j/p_i)^h\right\}\right]^{-1}<1.$$

From (18) and (19), it is clear that pricing decisions are now interdependent. We can represent each firm's pricing policy in terms of a price reaction function:

$$(20)\quad \begin{aligned} p_i &= F(p_j/p_i, \sigma, c)w \\ p_j &= F^*(p_i/p_j, \sigma, c)ew^*. \end{aligned}$$

Figure 3 shows the impact of a dollar appreciation in this setting. The schedules HH and H^*H^* are the reaction functions and A is the initial equilibrium.

An appreciation will shift the foreign reaction function up and to the left while leaving the home country's reaction function in place. The magnitude of the shift in H^*H^* at given relative prices (for example, along the OR ray) is proportional to the appreciation. Thus AB/AO represents the percentage appreciation. The new equilibrium is therefore at A'.

Note that this equilibrium at A' differs from the Dixit-Stiglitz one and resembles more nearly the Cournot model. Foreign firms reduce their price proportionally less than the reduction in dollar unit labor costs and home firms cut their price. But at A', the *relative* price of domestic products has increased compared to A as can be seen by the slope of a ray through A' compared to OR.

D. *Competition on the Circle*

I conclude the discussion of manufactures prices with a sketch of a third model of pricing for differentiated products. In the Dixit-Stiglitz model, consumers buy some of each brand of a product. Applied to toothpaste, that is an implausible model; we should therefore look for an alternative model where consumers buy only one brand. A particularly manageable version is the Salop model where consumers' tastes (defined by preferences for the attributes or characteristics of goods) are uniformly spread over the unit circle. Since domestic and foreign firms have potentially different costs, a symmetric equilibrium does not necessarily exist.

I simplify matters by assuming that there is an even number of firms, that domestic and foreign firms alternate along the circle, and that each consumer buys a unit from one or the other of the firms adjacent to his (her) preferred location. This is a very strong simplifying assumption because it implies a very different competition pattern from a circumstance where two foreign firms are adjacent.

Producers have constant unit labor costs and, other than entry costs, there are no fixed costs. With these assumptions, we can derive equilibrium prices and study the impact of dollar cost changes for foreign suppliers. Each consumer is located at a point on the circle. The significance of the location on the circle is that firms may not supply precisely the most preferred product. Accordingly, the consumer is forced to chose between the alternatives offered by the most adjacently located brands. Following Salop, consumers' surplus derived from buying a good that is a distance τ from the best location (on the circle) depends on the price and on the distance and the relationship is assumed linear:

$$h = v - c\tau - p, \tag{21}$$

where v is a constant, c denotes the utility cost per unit distance from the best location, and p is the price of a particular firm. Consumers will be indifferent between the brands offered by two competing firms on either side of their preferred location if the consumers' surplus is the same, $h_i = h_j$. Taking the case of n firms that are equally spaced on the circle, the condition for indifference between a domestic and a foreign supplier is

$$v - c\tau - p^* = v - c(1/n - \tau) - p. \tag{22}$$

Hence the distance served by a foreign firm is an increasing function of the price charged by domestic firms and a declining function of its own price:

$$\tau = (p + c/n - p^*)/2c. \tag{22'}$$

Profits for the foreign firm are equal to $2Lx$ times the excess of price over marginal cost:

$$\pi^* = (p^* - ew^*)2L(p + c/n - p^*)/2c, \tag{23}$$

where L denotes the total number of consumers and hence L also represents the density per unit distance served by the firm. Since the firm serves both sides of its location, $2L\tau$ is the total number of units sold. Maximization taking domestic price as given yields the foreign reaction function:

$$p^* = (p + c/n + ew^*)/2. \tag{24}$$

The typical domestic firm's reaction function is derived in the same manner:

$$p = (p^* + c/n + w)/2. \tag{25}$$

From (24) and (25), we obtain the solution for the prices charged by home and foreign firms:

$$p = c/n + (ew^* + 2w)/3; \qquad p^* = c/n + (2ew^* + w)/3. \tag{26}$$

From (26), we can calculate the elasticity of prices with respect to the exchange rate in this model:

$$\varphi = (1/3)(ew^*/p); \qquad \varphi^* = (2/3)(ew^*/p^*). \tag{27}$$

Note that these elasticities once again are fractions. If wages and hence prices are initially equal, $w = ep^*$, the elasticities simplify to the following expressions:

$$\varphi = \psi/3; \qquad \varphi^* = 2\psi/3; \qquad \psi \equiv 1/(1 + c/nw). \tag{27'}$$

The elasticities show that the relative price of imported goods declines and that the change in the relative price $\psi/3$ is smaller, the smaller the number of firms in the industry, and the lower the substitutability as measured by the term c. Along with the change in relative prices, there will be a shift in demand from home firms to foreign firms as consumers trade off the reduction in price for a larger distance from their most-preferred brand location.

At this point, it is worth commenting on the properties of the equilibrium when there is not an alternating pattern between domestic and foreign firms. Specifically, suppose that there are five firms, two domestic and three adjacent foreign firms. It is apparent that the middle foreign firm competes only with foreign firms, and hence will cut its price more than the outlying foreign firms that compete with home firms which have not experienced a cost reduction. Hence there will be three prices.

The same model can be applied to the foreign market. In terms of foreign exchange, the prices will rise and the relative price of our export brands abroad will rise. But, because it rises proportionally less than the currency appreciates, the export price in dollars changes in the proportion:

$$(28)\quad \varphi' = 1 - 2\psi^*/3;\ \ \psi^* = 1/(1 + c/n^*w),$$

where n^* is the total number of firms serving the foreign market. We can therefore find the change in the relative price of domestic exports in terms of imports and in terms of home brands:

$$(29)\qquad \varphi' - \varphi^* = 1 - 2(\psi + \psi^*)/3;$$
$$\varphi' - \varphi = 1 - 2\psi^*/3 - \psi/3.$$

The first point to note is that export prices may rise or fall in terms of import prices as a result of appreciation. But the fewer the number of firms in each country, the more likely that an appreciation leads to a *fall* in the relative price of exports. By contrast, as the number of firms increases (and hence ψ and ψ^* tend to unity), the relative price of exports *must* rise, reflecting the increase in the relative unit labor cost at home which sets competitive relative prices.

The second point is that export prices may decline relative to domestic prices as a result of an appreciation. This must be the case if the number of firms in the two markets is the same ($\psi = \psi^*$). As the number of firms increases, the relative price tends to remain unchanged. This result arises because price gets competed down to marginal cost which is the same for home and export production. It is apparent from (29) that the change in the relative price in terms of importables will always be larger than that in terms of domestic goods.

E. *Summary*

We have now seen common features of a number of models: they all predict that appreciation should lead to a decline in the price of imports. In the case of homogeneous goods, domestic firms, of course, fully match the decline in price. If products are differentiated, it will always be the case that the relative price of the imported brands declines in response to an appreciation. The extent of the decline depends on a measure of competition and on the relative number of home and foreign firms.

The empirically testable hypotheses concern price-marginal cost markups and the behavior of relative prices. For differentiated products, it is always the case that export and domestic prices will stay closer in line than import and domestic prices. In the Dixit-Stiglitz model, imports fall in terms of domestic goods and the relative price of exports goods stays unchanged in terms of home goods. In other models, the export price can in principle even decline in terms of imports.

IV. Some Evidence

Econometric testing of the hypotheses is unfortunately precluded by the absence of a comprehensive matched data set of export, import, and domestic prices. The BLS now publishes transactions prices for exports and imports that are disaggregated to the 4-digit level and classified on the SIC basis. But few of the series go back beyond the early 1980's. Where they do, the revisions of the SIC-based U.S. producer prices in most cases are either not all available yet, or only go back very few years. A complete overlap between export, import, and domestic prices for more than two years apparently only exists for fewer than a handful of cases, and overlap between any two series is limited to less than a dozen.

At a more informal level there are interesting patterns to observe. First, consider a comparison of U.S. export prices in dollars with those of Germany and Japan. Table 3 shows the percentage loss in U.S. competitiveness over the period 1980:IV to 1984:IV, using as a sample all available data at a highly disaggregated level. In the U.S.–Germany comparison, there are 36 different matched time-series, in the U.S.–Japan comparison, there are 20. Typical items in the list of commodities are "gears and gear units," or "household electrical space heating."

The data do not allow us to tell whether these are prices of the same products sold in the same third market (say France), or whether they represent exports to different markets (say U.S. sales to France and German sales in the United States). Accord-

TABLE 3—CHANGES IN RELATIVE PRICES: U.S. VS. GERMANY AND JAPAN
(Percentage change in relative export prices: 1980:IV–1984:IV)

	U.S.–Germany	U.S.–Japan
Mean	39.3	24.9
Standard Deviation	6.1	8.3

Source: See the Appendix.

TABLE 4—CUMULATIVE PRICE CHANGE: 1980:IV–1985:I

	Export Prices	Import Prices
Non-Electrical Machinery	18.0	−10.1
Scientific Instruments	18.0	−13.4

Source: See the Appendix.

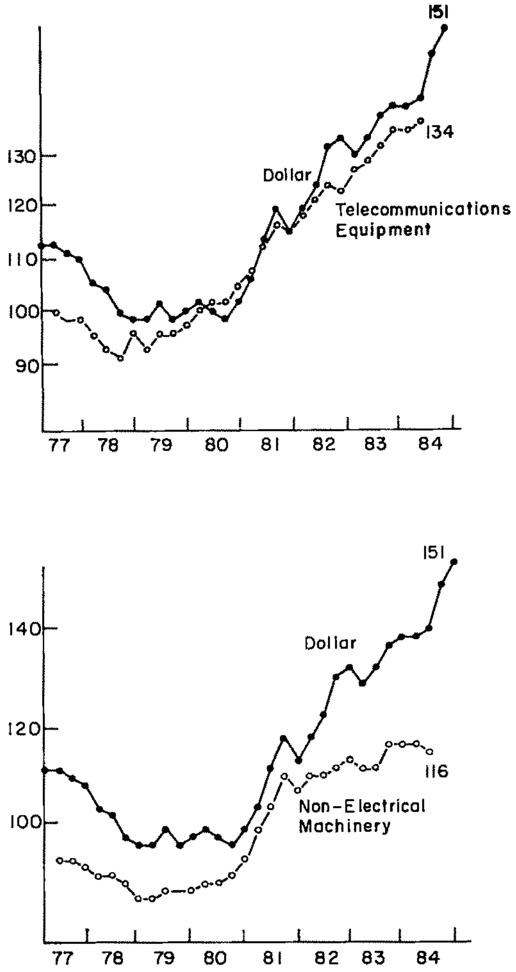


FIGURE 4

ingly, we cannot tell from these data whether they reflect market segmentation or imperfect substitution. They are consistent with markets being segmented, but goods being perfect substitutes and having a common price in the same market independent of supply source. But they are also consistent with markets being integrated—a common world market—but goods being imperfect substitutes so that the relative price of different suppliers can change.

Consider next a comparison of the transactions prices of U.S. exports and U.S. imports in the same commodity group. There is overwhelming evidence that, virtually without exception, the dollar appreciation of 1980–85 has been accompanied by an increase in the price of exports relative to imports. Evidence in this direction comes from export-import price comparisons at the more narrow 2- and 4-digit level. An example is provided in Table 4 which shows data for two 2-digit industries.

Figure 4 shows the ratio of export prices to import prices for telecommunications equipment and for nonelectrical machinery. The figure also show the index of the *nominal* dollar exchange rate index. The dollar appreciation since 1980 gives rise to an increase in the relative price of exports in terms of imports. Table 4 shows indices of the relative export-import price for all series where comparable SIC data exist. The same pattern would be obtained by comparing U.S. to German and Japanese export prices in these individual commodity groups. The first finding then is, that across industries, virtually without exception, export prices have increased relative to import prices. This is true at the level of individual commodities, but also, as shown at the outset of the paper, for aggregate export and import unit values.

This result would obtain strictly only in the Dixit-Stiglitz model. In the other formulations, it is a possibility though it need not occur. Tables 5–7 look at the price of exports and imports relative to each other and

TABLE 5—THE RATIO EXPORT TO IMPORT PRICES
(Index 1980:I = 100)

	2011	301	35	353	356	3569	357	3643	38
1979:IV	105	103	–	100	95	96	–	91	92
1981:IV	108	105	112	118	119	121	106	115	108
1985:I	126	104	131	135	152	143	110	152	136

Note: The headings are SIC codes. For definitions and source, see the Appendix.

TABLE 6—THE RATIO OF EXPORT TO DOMESTIC PRICES
(Index 1980:IV = 100)

	3546	3555	3674	3533	3523	3519	3494	2011	3537
1979:IV	101	101	109	99	99	103	99	110	97
1981:IV	100	104	91	100	100	103	103	93	99
1985:I	95	107	93	100	102	105	108	108	100

Note: See Table 5.

TABLE 7—THE RATIO OF IMPORT TO DOMESTIC PRICES
(Index 1980:IV = 100)

	2311	2033	3651	3143	3531	2435	2011	3312	3313
1979:IV	100	108	–	96	98	120	105	101	88
1981:IV	101	92	100	95	90	114	92	98	88
1985:I	110	90	92	88	76	102	85	84	74

Note: See Table 5.

relative to domestic producer prices in the commodity group. Export prices change little relative to domestic prices, even though there is no clear pattern of decline in all industries. By contrast, most import prices decline in terms of domestic goods. But the order of magnitude of the decline remains relatively small compared to the change in relative unit labor costs. With a change in relative unit labor costs of more than 40 percent, the decline in the relative price is in most cases less than 20 percent. That is not at all out of line with the theory once some degree of "pricing to the American market" is taken into account, just as the price-setting models above suggest.

It is worth noting that, at the retail level, this effect would obtain even more strongly. The reason is that here distribution costs come into play, so that even with the full pass-through of cost reductions on imported goods, the proportional decline in the price of imported goods would be much less than the exchange rate appreciation.

V. Concluding Remarks

The models reviewed in this paper focus on a relatively short time perspective. The wage rate is assumed not to react to changes in output and profitability, and the number and location of firms in an industry is unaffected. These assumptions are plausible in the short term, but it is clear that a sustained real appreciation will ultimately show its effects in wage cuts in those industries where the loss in competitiveness causes unemployment and wage increases in the expanding sectors. Firms will close in high-wage areas and entry into an industry will take place in areas where labor costs are low. These longer-term adjustments are also part of the macroeconomics of adjustment to exchange rate movements. They imply that the absolute and relative numbers of firms (n and n^*) will be endogeneous as the location of firms on the product circle.

It is clear from the analysis offered here that for these issues, a microeconomic perspective will also be helpful. In particular, it will be interesting to see how pricing decisions are affected by entry and relocation possibilities at an international level, and by the anticipated persistence of disequilibrium exchange rates.

The analysis developed here has application not only to the exchange rate question, but also to the short-term effect of trade liberalization. The common argument is that a small country by opening up can take advantage of the world markets, enjoying price reductions in proportion to the tariff reduction. That clearly assumes perfect competition. If markets are less than fully competitive, the analysis offered here becomes relevant to the trade liberalization issue.

DATA APPENDIX

The data in Table 2 are unit values obtained from Data Resources, Inc., U.S. Central Data Bank.

The data in Table 3 are compiled by the Bureau of Labor Statistics, Division of International Prices. The unpublished data are from a compilation entitled "Comparisons of U.S., German and Japanese Export Price Indices," dated May 1985.

The data for export and import prices in Tables 4–7 are SIC-based price data compiled by the Branch of Export and Import Price Indexes of the Bureau of Labor Statistics. The data are from a printout, dated July 1985. The domestic price index for SIC-based producer prices is obtained from Data Resources, Inc. Table 4: The data refer to SIC codes 34 and 38, respectively. In Tables 5–7, the following are the definitions of the SIC codes shown in the table heading:

Table 5: 2011-Meat and Meat Packing Products; 301-Tires and Inner Tubes; 35-Machinery, except electrical; 353-Construction, Mining Equipment; 3569-General Industrial Machines; 357-Office, Computing and Accounting Machinery; 3643-Current-Carrying Wiring Devices; 38-Scientific Instruments, Optical Goods, Clocks.

Table 6: 3546-Power-Driven Hand Tools; 3555-Printing Trades Machines, Parts; 3674-Semiconductor Devices; 3533-Oil and Gas Field Equipment; 3523-Farm Machinery and Equipment; 3519-Internal Combustion Engines; 3494-Metal Valves, Pipe Fittings; 2011-Meat and Meat-Packing Products; 3537-Industrial Trucks, Portable Elevators.

Table 7: 2311-Men's or Boy's Suits & Coats, except Raincoats; 2033-Canned and Preserved Fruits, Vegetables, Jams, Juices; 3651-Video and Audio Equipment; 3143-Men's Footwear except Athletic; 3531-Construction Machinery; 2435-Hardwood Veneer and Plywood; 2011-Meat and Meat-Packing Products; 3311-Rolling Mill Products; 3313-Electrometallurgical Products.

The index of the nominal dollar exchange rate in Figure 4 is the Morgan Guaranty trade weighted index of the dollar. The series was obtained from Data Resources, Inc. The series for telecommunications equipment and non-electrical machinery are unpublished, SITC-based prices of exports and imports obtained from the Division of International Prices, Bureau of Labor Statistics.

REFERENCES

Aizenman, Joshua, "Monopolistic Competition and Deviations from *PPP*," unpublished manuscript, University of Chicago, 1985.

______, "Testing Deviations from Purchasing Power Parity (*PPP*)," *Journal of International Money and Finance*, March 1986, *5*, 25–35.

Blanchard, Olivier, "Monopolistic Competition, Small Menu Costs, and Real Effects of Nominal Money," unpublished manuscript, MIT, 1985.

Dixit, Avinash, "International Trade Policy for Oligopolistic Industries," *Economic Journal*, Suppl. 1984, *94*, 1–16.

______, "Comparative Statics for Oligopoly," *International Economic Review*, February 1986, *27*, 107–122.

______ **and Stiglitz, Joseph,** "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Dornbusch, Rudiger, "Flexible Exchange Rates and Interdependence," *IMF Staff Papers*, March 1983, reprinted in *Dollars, Debts and Deficits*, Cambridge: MIT Press, 1986.

______, "Purchasing Power Parity," NBER Working Paper No. 1591, 1985, in *Palgrave's Dictionary of Economics*, London: Macmillan, forthcoming.

______, "Inflation, Exchange Rates and Stabilization," *Essays in International Finance*, International Finance Section, Princeton University, 1986.

Giovannini, Alberto, "Exchange Rates and Traded Goods Prices," unpublished manuscript, Columbia University, 1985.

Feinberg, Robert M., "The Interaction of Foreign Exchange and Market Power Effects on German Domestic Prices," *Journal of Industrial Economics*, September 1986, 61–70.

Flood, Eugene, Jr., "An Empirical Analysis of the Effects of Exchange Rate Changes on Goods Prices," unpublished manuscript, Stanford University, 1986.

Hart, Oliver, "A Model of Imperfect Competition with Keynesian Features," *Quarterly Journal of Economics*, February 1982, *97*, 109–38.

Helpman, Ephanan, and Krugman, Paul, *Market Structure and Foreign Trade*, Cambridge: MIT Press, 1985.

Kravis, Irving and Lipsey, Robert, "Price Behavior in the Light of Balance of Payments Theories," *Journal of International Economics*, May 1978, *8*, 193–246.

______ and ______, "Prices and Terms of Trade for Developed Country Exports of Manufactured Goods," in B. Csikos-Nagy et al., eds., *The Economics of Relative Prices*, New York: Saint Martin's Press, 1984.

Krugman, Paul, "Pricing to Market when the Exchange Rate Changes," NBER Working Paper No. 1926, May 1986.

Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly," *Quarterly Journal of Economics*, May 1985, *100*, 529–37.

Mann, Catherine, "Prices, Profit Margins and Exchange Rates." *Federal Reserve Bulletin*, June 1986, *72*, 366–79.

Seade, Jesus, "Prices, Profits and Taxes in Oligopoly," Working Paper, University of Warwick, 1983.

Salop, Steven, "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, Spring 1979, *10*, 141–56.

Exchange Rate Management: Intertemporal Tradeoffs

By ELHANAN HELPMAN AND ASSAF RAZIN*

Exchange rate management is possible only if the government pursues consistent monetary and fiscal policies. We construct a model in which the real consequences of exchange rate management depend on the precise time pattern of these policies. We study the constraints on feasible policies and the comparative dynamics of disinflation by means of exchange rate targetting. Our theoretical results are consistent with exchange rate-managed disinflation attempts in Argentina, Chile, and Israel.

It is now understood that exchange rates cannot be managed without the pursuit of other policies which make the entire package internally consistent (see, for example, J. J. Polak, 1957). Governments or central banks can only temporarily target exchange rates without giving due attention to other policies. If they do, they have to choose (or are forced to choose) eventually measures which validate *ex post* the feasibility of their exchange rate policy. These measures will typically be anticipated by economic agents, thereby generating immediate pressures in various markets. Hence, the success of exchange rate management depends to a large extent on other policies, on commitments to future policies, and on their effects on expectations.

A major purpose of this paper is to study the effects of policy-induced slowdowns in the rate of currency devaluation, which are not accompanied by an immediate monetary or fiscal contraction that prevents reserve losses, thereby implying the need for a contractionary policy in future periods.[1] An extreme form of this policy is an exchange rate freeze. Our interest in experiments of this type stems from the fact that several countries have attempted to disinflate by means of slowdowns in the rate of currency depreciation, with Argentina, Chile, and Israel being the prime examples. Argentina used a preannounced pattern of exchange rate movements (Tablita) from December 1978 to February 1981. Chile used a Tablita from February 1978, which culminated in an exchange rate freeze in June 1979. The frozen exchange rate was maintained until June 1982. Israel used a preannounced rate of currency devaluation from September 1982 until October 1983 (at a rate of 5 percent per month).[2] In all cases, the currency was overvalued and the managed rate of currency devaluation was below the inflation rate. Table 1 presents data for these countries. It is clear from these data that in all cases, the policy brought about an appreciation in the real exchange rate, an increase in private consumption, and a worsening of the trade balance. In all cases the exchange rate management policy turned out to be unsustainable.

In order to study these issues we construct a model of overlapping generations in which

*Department of Economics, Tel Aviv University, Ramat Aviv, Tel Aviv 69 978, Israel. This is an expanded version of NBER Working Paper No. 1590. We thank the Horowitz Institute and the Foerder Institute for Economic Research, Tel Aviv University, for financial aid, and participants at the Tel Aviv Workshop on International Economics, the Maale Hachamisha conference on Exchange Rate Management, the Paris Macroeconomic seminar, Western Ontario Money Workshop, the New University of Lisbon seminar, and the 1984 NBER Summer Institute, for comments.

[1] Since the appearance of the first version of this paper, other studies have also dealt with related issues, including Maurice Obstfeld (1985, 1986) and Sweder van Wijnbergen (1986). However, those studies emphasized different points.

[2] An exchange rate freeze has also been part of the Argentinian and Israeli stabilization programs of mid-1985, but they were different from their predecessors in other important ways. We return to this point in the closing section.

TABLE 1

	Index of Real Exchange Rate[a]	Index of Real Private Consumption[b]	Trade Balance (millions of U.S. dollars)[c]
Argentina			
1978	100	100	2,913
1979	73	111	1,782
1980	72	109	−1,373
1981	85	107	712
Chile			
1978	100	100	−426
1979	88	119	−355
1980	75	122	−764
1981	–	132	−2,677
Israel			
1981	100	100	−1,371
1982[d]	96	106	−1,837
1983	91	115	−2,741
1984	98	110	−2,130

[a] *Sources:* Argentina—Rudiger Dornbusch (1985); Chile—Arnold Harberger (1982); Israel—Israel Central Bureau of Statistics, *National Accounts 1972–1985*, Appendix to *Israel Statistical Monthly*, No. 5, 1986. The Index for Israel was computed by means of the implicit price deflators of imports and GDP. These numbers are for calendar years. Therefore, for Israel in 1982, the number represents an underestimation of the year 1981:IV–1982:III.

[b] *Source* for Argentina and Chile is *International Financial Statistics*, December 1984, vol. 37. Argentina and Chile were calculated as row 96f divided by row 64.

[c] Israel's trade balance is the current account balance minus interest payments minus defense imports. Argentina and Chile were calculated as row 77aad minus row 77abd.

[d] Except for the index of the real exchange rate, length of year defined as quarter of previous year plus the first three quarters of the current year. This is done due to the fact that exchange rate management began in the last quarter of 1982.

consumers have finitely expected horizons, as in Olivier Blanchard (1985). This model is particularly suitable for the purpose at hand because it does not have the Ricardo-Barro neutrality property (see Robert Barro, 1974). It is well known that in economies without distortions, in which individuals have infinite horizons, exchange rate management has no real effects (for example, see our earlier paper, 1979). Therefore, these types of models cannot explain the facts depicted in Table 1. On the other hand, the existence of distortions, or finite horizons, introduces the possibility of real effects of exchange rate management (for examples of real effects that result from distortions, see David Aschauer and Jeremy Greenwood, 1983; our paper, 1984; and Robert Feenstra, 1986). We choose, however, to rely on the presence of finite horizons in order to explain the macroeconomic performance reported in Table 1.

We present our model in Section II and discuss a benchmark equilibrium with a freely floating exchange rate. The benchmark is chosen in order to highlight the effects of exchange rate management. In that framework, the time pattern of real variables does not depend on monetary injections or withdrawals through the tax-transfer system, and the resulting equilibrium is identical with the equilibrium that would have resulted in a barter economy. This leads to an efficient allocation of resources. Efficiency is also preserved when the exchange rate is managed, but the time pattern of real variables *does* depend on the policies that support the exchange rate path (in contrast with our 1979 paper, Helpman, 1981, and Robert Lucas, 1982). Hence, in this case there exist significant differences between a managed and a floating exchange rate regime. In Section III we study these issues in general terms. In particular, we explore the feasibility of various exchange rate policies in conjunction with the accompanying fiscal and monetary policy mix. The real effects of exchange rate targetting that arise in this analysis are closely related to the real effects of budget financing that were discussed by Blanchard, and by Jacob Frenkel and Razin (1986) in a framework without money. Here, we naturally take explicit account of monetary considerations.

In Section IV we study in detail the extreme form of exchange rate management—the case of an exchange rate freeze. This is designed to shed light on the economic mechanism underlying disinflation policies by means of exchange rate targetting. We show that when this policy is pursued with an initially overvalued currency (excessively low nominal exchange rate) and a delayed absorption policy, the result is higher spending and a low real exchange rate following the inception of exchange rate management, lower spending and higher real exchange rates in later periods, and larger aggregate external debt in all time periods. The twist in the time profile of spending and the real exchange rate, and the upward shift in the time profile of debt, are more pronounced the larger the initial overvaluation and the longer the delay in the absorption policy.[3] However, the delay in the absorption policy is bounded by the government's taxing capacity. Therefore, the feasible delay in the absorption policy is also bounded. The beneficiaries of this policy combination are individuals who are alive during its inception, while all future generations suffer.

I. Floating Exchange Rate

We consider an economy with overlapping generations in which a cohort of size 1 is born in every period. Individuals survive to the next period with probability γ and this probability is age independent. The event of death is independent across individuals. Therefore, the proportion of a cohort alive at time t_1 which survives to period t_2 is $\gamma^{(t_2-t_1)}$.

The age distribution of the population is constant over time and in every period there are γ^a individuals of age a. The size of the population is also constant and equal to

$$\sum_{a=0}^{\infty} \gamma^a = 1/(1-\gamma).$$

We assume that those individuals live in a small country facing a given one-period world real interest rate r on sure loans in terms of traded goods. All loans are indexed and foreign prices of traded goods are constant and equal to one. Thus, if one borrows b, he (she) has to repay Rb the next period, where $R=1+r$ is the interest factor. Since an individual survives to the next period only with probability γ, he cannot obtain a loan with this interest rate. Foreign financial institutions who lend to domestic residents will obtain a sure repayment Rb if they charge a real interest rate of $(R/\gamma)-1$. In order to see this, suppose that b is being lent to every individual of a given cohort. Then, those who will survive to the next period will repay Rb/γ. However, only a proportion γ of the cohort will survive. Therefore total payments by the cohort will be Rb. Clearly, $(R/\gamma)-1$ is the risk-adjusted real interest rate (see Blanchard).

There exist firms that produce y_T units of traded goods per capita and y_N units of nontraded goods per capita. The sectoral output levels are functions of the relative price of nontradables p_t, and so is gross domestic product per capita in terms of traded goods, which we denote by $y(p_t)$.

Firms sell their output in exchange for domestic money and distribute the proceeds to the living individuals at the beginning of the following period. Individuals have to pay for goods with money; with home money for home goods and foreign money for foreign goods. Thus, we assume a system with cash-in-advance constraints in which goods are bought with the seller's currency (see our 1984 paper for a discussion of alternative monetary mechanisms). A firm pays its owners the proceeds from period $t-1$ at the beginning of period t. At the beginning of

[3] See also Allan Drazen and Helpman (1986) for a discussion of the effects of uncertainty in the timing of the absorption policy.

period t, every individual receives also the repayment (inclusive of interest) of loans he gave in $t-1$, and new loans are issued. These transactions result in a stock of money that is allocated between domestic and foreign money by trading in the foreign exchange market. These final stocks of money are then used during period t to purchase goods. Local firms absorb during period t the entire stock of domestic money and pay it out as dividends at the beginning of period $t+1$. This sequence of transactions is repeated in every period.

Assuming that the price of tradeables in terms of foreign currency is equal to one, the budget constraint of an individual of age a in period t is

$$\text{(1a)}\quad c_{a,t}=b_{a,t}-(R/\gamma)b_{a-1,t-1}+\left[e_{t-1}y(p_{t-1})-\tilde{\theta}_t\right]/e_t,$$

where $c_{a,t}$ $(\equiv c_{Ta,t}+p_t c_{Na,t})$ is his total consumption in terms of traded goods ($c_{Ta,t}$ is his consumption of traded goods and $c_{Na,t}$ is his consumption of nontraded goods); $b_{a,t}$ is his new debt; $(R/\gamma)b_{a-1,t-1}$ is his repayment of old debts, with period minus-one debt equal to zero (i.e., $b_{-1,t}=0$ for all t); e_t is the exchange rate; and $\tilde{\theta}_t$ are the age-independent nominal taxes or transfers. The last term on the right-hand side of (1a) represents the foreign currency value of period t income from firms (that consists of their period $t-1$ sales) minus tax payments. We also assume in (1a) that money is not used for store of value purposes, which is guaranteed when the nominal effective interest rate is positive. In addition we need to impose the solvency (terminal) condition:

$$\text{(1b)}\quad \lim_{\tau\to\infty}(R/\gamma)^{-\tau}b_{a+\tau,t+\tau}=0.$$

We assume that individuals maximize expected lifetime utility:

$$\text{(1c)}\quad E_t\sum_{\tau=0}^{\infty}\delta^{\tau}v(p_{t+\tau},c_{a+\tau,t+\tau})=\sum_{\tau=0}^{\infty}(\gamma\delta)^{\tau}v(p_{t+\tau},c_{a+\tau,t+\tau}),$$

subject to (1a) and (1b), where $v(\cdot)$ is the temporal indirect utility function and δ is the subjective discount factor.

We assume that the government has no real spending. In a freely floating exchange rate regime, the government injects and withdraws money from the economy via the taxes and transfers $\tilde{\theta}_t$. Hence, if m_t is the per capita stock of money in period t, then

$$\text{(2)}\quad m_t=m_{t-1}-\tilde{\theta}_t.$$

On the other hand, with positive nominal interest rates all domestic money is spent on domestic goods (see Helpman), implying $m_t=e_t y(p_t)$. Taken together, the last two equations imply

$$\text{(3)}\quad \left[e_{t-1}y(p_{t-1})-\tilde{\theta}_t\right]/e_t=y(p_t).$$

Using (1) and (3) it is clear that in a pure floating exchange rate regime, monetary injections and withdrawals via $\tilde{\theta}_t$ have no real effects. Put differently, in this system, the time pattern of consumption and real debt (individual as well as aggregate), and the real exchange rate ($1/p_t$) (i.e., the price of tradeables in terms of nontradeables), do not depend on the time pattern of monetary injections. This result is in line with models in which there are no overlapping generations and individuals live to the end of the economy's horizon (see Helpman).

We use this case as a benchmark of comparison with exchange rate management policies. It should be pointed out, though, that neutrality of the above described monetary policy need not hold in the presence of a labor-leisure choice or externally financed investment. It also need not hold if the buyer's currency is used for transactions instead of the seller's currency (see our 1984 paper). Moreover, a special feature of the current formulation of overlapping generations is that the neutrality result also depends on the assumption of an equal division of period $t-1$ proceeds among all living individuals at the beginning of period t. It can, for example, be shown that if this is not so, and individuals can take loans using current period output as collateral (in case they do not survive to the next period), then monetary policy will have real effects. We do not pursue this line in order to avoid sidetracking.

II. Intertemporal Constraints on Exchange Rate Management

Suppose that in the benchmark case considered in the previous section, we obtain the following solution for per capita consumption and debt:

$$\bar{c}_t = (1-\gamma)\sum_{a=0}^{\infty}\gamma^a\bar{c}_{a,t},$$

$$\bar{b}_t = (1-\gamma)\sum_{a=0}^{\infty}\gamma^a\bar{b}_{a,t}, \qquad t=-\infty,\dots,$$

and the real exchange rate $\{1/\bar{p}_t\}_{t=-\infty}^{\infty}$. Suppose also that in period $t=0$, the government begins to manage the exchange rate. The question we consider is: what are the real consequences of exchange rate management in terms of deviations of $\{c_t, b_t, p_t\}$ from the benchmark case? As we know from Helpman, in a model without overlapping generations in which the individual's life extends to the economy's horizon, exchange rate management has no real effect as long as the government is intertemporally balanced, independently of the time pattern of taxes. The reason is that, in that case, the private sector fully internalizes the government's intertemporal budget constraint. This, however, cannot be expected in an economy in which individuals pay future taxes with a probability smaller than one. Therefore, the time pattern of taxes required in order to manage the exchange rate will generally have real effects.

Clearly, for every time pattern of exchange rates there exists a time pattern of taxes and transfers which preserves the benchmark real variables. However, while in previous models (for example, Helpman) there was no constraint on the time pattern of the neutral taxes but only on their present value, here this pattern is *unique*. As is clear from (1a) and (3), if in period $t=0$ the real exchange rate remains $(1/\bar{p}_0)$ and per capita debt does not change, then for a given pattern of exchange rates $\{e'_t\}_{t=0}^{\infty}$ the taxes $\{\tilde{\theta}'_t\}_{t=1}^{\infty}$ have to satisfy

$$(4) \qquad \tilde{\theta}'_t = e'_{t-1}y(\bar{p}_{t-1}) - e'_t y(\bar{p}_t), \qquad t=0,1,\dots,$$

where $e'_{-1} = \bar{e}_{-1}$. This policy generates monetary injections and withdrawals which keep the money supply in line with the nominal value of output implied by the exchange rate e'_t and the real exchange rate $(1/\bar{p}_t)$; that is, it ensures

$$m'_t = m'_{t-1} - \tilde{\theta}'_t = e'_t y(\bar{p}_t)$$

with no deficits or surpluses in the overall balance of payments. Thus, for example, if the next-period-managed exchange rate brings about a decline in the demand for money, the monetary contraction via the tax system matches this decline when taxes are chosen according to (4). If (4) is satisfied, we have

$$(5) \qquad [e'_{t-1}y(\bar{p}_{t-1}) - \tilde{\theta}'_t]/e'_t = y(\bar{p}_t), \qquad t=0,1,\dots,$$

which is the same as (3), implying no change in real variables. In this case, no reserve movements are required in order to manage the exchange rate. Observe, however, that even when the exchange rate is maintained constant over time, *varying* taxes and transfers are required as long as the real exchange rate is not constant over time. Hence, *exchange rate management without real consequences requires a well-coordinated time-varying policy of monetary injections and withdrawals*. No such policy is required under a free float.

It is clear from this discussion that if a policy that satisfies (3) does not accompany the exchange rate management programme, there will be reserve movements. It is assumed that reserves are bearing interest. In this case, reserve movements generate public debt which has real effects. The time pattern of the public debt per capita is given by

$$(6a) \qquad b^G_t = Rb^G_{t-1} + \frac{1}{e'_t}[e'_{t-1}y(p'_{t-1}) - e'_t y(p'_t) - \tilde{\theta}'_t], \qquad t=0,1,\dots,$$

with the initial conditions:

$$\text{(6b)} \quad b^G_{-1}=0,\; e'_{-1}=\bar{e}_{-1},\; p'_{-1}=\bar{p}_{-1}.$$

Equation (6a) describes reserve (external interest-bearing assets) movements according to the standard balance-of-payments mechanism. External debt grows at the rate of interest due to rollovers, plus periodical additions through deficits in the overall balance of payments. The last component is represented by the terms in the square brackets. The first two terms describe the decline in the overall demand for money $[e'_{t-1}y(p'_{t-1})-e'_t y(p'_t)]$. Part of this decline is satisfied by negative injections (withdrawals) via taxes $\tilde{\theta}'_t$. The rest is attained via foreign exchange purchases by the private sector, which bring about reserve losses. It is clear from (6a) that when (4) is satisfied, we obtain the solution $p'_t=\bar{p}_t$ and $b^G_t=0$ for all t; that is, there are no reserve movements.

We assume that the government repays its debts. Therefore, its policy is restricted to satisfy (6a) and (6b) with the terminal condition:

$$\text{(6c)} \qquad \lim_{t\to\infty} R^{-t}b^G_t=0.$$

Now, suppose that the exchange rate management policy starts in period zero and the policy rule given in (4) is not followed. Then the effects of reserve movements on individual budget constraints can be seen from the following rewriting of the budget constraint (1a), using (6a):

$$\text{(7)} \quad c'_{a,t}=b'_{a,t}-(R/\gamma)b'_{a-1,t-1} + y(p'_t)+\left(b^G_t-Rb^G_{t-1}\right).$$

In order to see as clearly as possible the real effects of reserve movements, assume for the remaining part of this section that all goods are traded; that is, $y_N=0$ and $y(p_t)\equiv y_T$. Then (1a), (1b), and (3) imply (for the flexible exchange rate case):

$$\text{(8)} \quad \sum_{\tau=0}^{\infty}(\gamma/R)^\tau c_{a+\tau,t+\tau} = \sum_{\tau=0}^{\infty}(\gamma/R)^\tau y_T-(R/\gamma)b_{a-1,t-1},$$

while (7) and (1b) imply (for the managed exchange rate case):

$$\text{(9)} \qquad \sum_{\tau=0}^{\infty}(\gamma/R)^\tau c'_{a+\tau,t+\tau}=w'_{a,t},$$

where $w'_{a,t}$ is real wealth of an individual of age a, defined as

$$\text{(10)} \qquad w'_{a,t}=\sum_{\tau=0}^{\infty}(\gamma/R)^\tau y_T -(R/\gamma)b'_{a-1,t-1}-Rb^G_{t-1} +(1-\gamma)\sum_{\tau=0}^{\infty}(\gamma/R)^\tau b^G_{t+\tau}.$$

It is clear from a comparison of (8) with (9) and (10) that reserve movements have real effects and that these effects depend on the *time pattern* of reserve movements. Different time patterns of reserves generate different redistributions of wealth across generations, thereby effecting the time pattern of aggregate spending, as we will explicitly show in the next section. Observe also that if we assume that no new cohorts are born and the probability of survival equals one, then viewed from $t=0$, constraints (9) and (10) coincide with (8), which implies neutrality of the exchange rate policy.

The redistributional effects embodied in (10) can also be seen in another way, by combining it with (6) in order to obtain

$$\text{(11)} \quad w'_{a,t}=\sum_{\tau=0}^{\infty}(\gamma/R)^\tau\left(y_T-\frac{\tilde{\theta}'_{t+\tau}}{e'_{t+\tau}}\right) +\sum_{\tau=0}^{\infty}(\gamma/R)^\tau\left(\frac{e'_{t+\tau-1}-e'_{t+\tau}}{e'_{t+\tau}}\right)y_T -(R/\gamma)b'_{a-1,t-1}.$$

It is seen from here that the real wealth of an individual born at time t depends both on the given path of exchange rate depreciation rates and on the path of taxes. An individual born at t is better off the further away in the future taxes are imposed and the exchange rate is depreciated, and the nearer in the future transfers are given and the exchange rate is appreciated. However, the taxes cannot be divorced from exchange rate movements, because from (6) they have to satisfy

$$(12)\quad \sum_{t=0}^{\infty} R^{-t}(\tilde{\theta}'_t/e'_t) = \sum_{t=0}^{\infty} R^{-t}\frac{1}{e'_t}[e'_{t-1}y(p'_{t-1})-e'_t y(p'_t)],$$

where $p'_{-1}=\bar{p}_{-1}$ and $e'_{-1}=\bar{e}_{-1}$, with $y(p_t) \equiv y_T$ in the case of traded goods only. This equation states that the present value of taxes should equal the present value of changes in the demand for money, or that the present value of reserve losses is zero.[4] Hence, the larger are the rates of currency appreciation in the present and the near future (close to $t=0$), the heavier should be future tax burdens or the larger should be future currency depreciations.

Coming back to (11), observe that from period $t=1$ onwards, exchange rates and taxes are fully anticipated. However, at time $t=0$, when exchange rate management begins, there is an *unanticipated* change in both the exchange rate and in taxes. It is therefore useful to decompose the contribution of period-zero disposable income, inclusive of capital gains on wealth, into anticipated and unanticipated components. Given the discussion of the benchmark case, it is clear that the anticipated component of real income is y_T, while from (11) total real income is

$$y_T-\frac{\tilde{\theta}'_0}{e'_0}+\frac{\bar{e}_{-1}-e'_0}{e'_0}y_T.$$

Therefore, the last two terms represent the unanticipated components, and they can be expressed as

$$(13)\quad -\frac{\tilde{\theta}'_0}{e'_0}+\frac{\bar{e}_{-1}-e'_0}{e'_0}y_T=g=k-h,$$

where (since $\bar{e}_0=\bar{e}_{-1}y_T-\tilde{\theta}_0$)

$$(13')\quad k=[(1/e'_0)-(1/\bar{e}_0)]\bar{e}_{-1}y_T$$

is the unanticipated capital gain on money balances and

$$(13'')\quad h=(\tilde{\theta}'_0/e'_0)-(\tilde{\theta}_0/\bar{e}_0)$$

is the unanticipated increase in real taxes.[5]

The definition of the real loss due to taxes is clear from (13″), while the capital gain on money balances may require an explanation. Before the change in the exchange rate policy, money balances held by the private sector are $\bar{m}_{-1}=\bar{e}_{-1}y_T$. The real value of this money was expected to be $y_T\bar{e}_{-1}/\bar{e}_0$. As a result of the unanticipated stabilization of the exchange rate at e'_0, the real value of this money has become $\bar{e}_{-1}y_T/e'_0$. Hence, (13′) describes the unexpected capital gain on period-minus-one money holdings.

Now, using (13) it is seen from (6a) (with $y(p)=y_T$) that

$$(14)\quad b_0^G=g=k-h.$$

Namely, the initial loss of reserves is equal to the public's unanticipated capital gain on money holdings minus the unanticipated increase in tax obligations. In view of (14), condition (4) for $t=0$ (which describes the period-zero absorption policy that is required for real neutrality of the exchange rate management policy) can be interpreted as follows: The tax rate $\tilde{\theta}'_0$ is chosen so as to make the unanticipated increase in tax liabilities h just equal to the unanticipated capital gain on money balances k. When this holds, there are no initial reserve movements.

III. Disinflation by Means of an Exchange Rate Freeze

In order to evaluate the macroeconomic effects of exchange rate management, we consider in this section the extreme case of

[4] Equation (12) can also be interpreted as saying that the present value of lump sum taxes plus inflation taxes is equal to zero. This is the proper constraint for a government which purchases no goods and services and starts with no debt.

[5] In the presence of nontraded goods, there exists also a capital gain due to the unanticipated movement in the real exchange rate.

an exchange rate freeze. For concreteness, suppose that the government freezes the exchange rate at the level e from period zero to infinity. In this case, real effects are prevented if taxes equal $\tilde{\theta}_t'$, $t=1,2,\dots$, as given in (4). As shown in the previous section, real neutrality requires the taxes to vary over time in reaction to real exchange rate movements. When other real world complications are added as well, such as the dependence of the velocity of circulation on the nominal interest rate, the required policy coordination for real neutrality becomes even more complicated.

We choose to analyze the following policy experiment. Suppose that the exchange rate e is chosen below $\bar{e}_0$ (below the equilibrium exchange rate in a freely floating system) and taxes are not changed in period zero. Starting with period $t=1$, taxes are imposed according to (4) until period $t=T-1$. This level of taxation is too low to prevent reserve losses. From $t=T$ onwards, fixed real taxes $\tilde{\theta}'/e$ are added to the taxes in (4) in order to pay interest on public foreign debt, so that the government budget is balanced for $t=T$, $T+1,\dots$. In this case, exchange rate management begins with an overvalued currency; we wish to explore the resulting real effects and their dependence on the timing of the contractionary policy.

We have selected a particular tax structure to accompany the exchange rate freeze which is both feasible (it satisfies (12)) and simple to understand. Other variants can, of course, also be considered, but it seems to us that the simplicity of the tax structure is an advantage as far as the current illustration is concerned. Our scheme implies that in every period $t\in\{1,2,\dots,T-1\}$, taxes are just equal to the decline in the demand for money (see (4)). In this case, the consolidated budget deficit (of the government and the central bank) is equal to interest payments on the public debt. Hence, from $t=1$ to $t=T-1$, public debt grows at a rate that equals the rate of interest $R-1$. From $t=T$ onwards, taxes are equal to the decline in the demand for money plus a constant that covers interest payments on the public debt. The result is that from $t=T-1$ onwards, public debt remains constant.

Let $g=b_0^G>0$ be the initial public debt that results from the exchange rate freeze. Then under our tax scheme, (6a) implies

$$(15)\quad b_t^G=\begin{cases} R^t g, & 0\le t\le T-1\\ R^{T-1}g, & t\ge T\end{cases}$$

That is, the government's debt grows at the rate of interest until period $T-1$ and remains constant afterwards.

Using (4) and (15), taxes are given by

$$\tilde{\theta}_t'/e=\begin{cases} y(p_{t-1})-y(p_t) & \text{for } 1\le t\le T-1\\ y(p_{t-1})-y(p_t) \\ \quad +(R-1)R^{T-1}g & \text{for } t\ge T\end{cases}$$

In this case, the individual budget constraint (1a) can be written as

$$(16)\quad b_{a,t}=(R/\gamma)b_{a-1,t-1} + c_{a,t}-[y(p_t)-\theta_t],$$

where the effective tax rates θ_t (including capital gains) are given by

$$(17)\quad \theta_t=\begin{cases} -g & t=0\\ 0 & 1\le t\le T-1\\ (R-1)R^{T-1}g & t\ge T\end{cases}$$

Now, (16) implies that the individual budget constraint is

$$(18)\quad \sum_{\tau=0}^{\infty}(\gamma/R)^{\tau}c_{a+\tau,t+\tau} = \sum_{\tau=0}^{\infty}(\gamma/R)^{\tau}[y(p_{t+\tau})-\theta_{t+\tau}] -(R/\gamma)b_{a-1,t-1}\equiv w_{a,t}.$$

Individuals maximize expected lifetime utility (1c) subject to (18), which replaces (1a) and (1b). For current purposes, assume also that the temporal indirect utility function $v(\cdot)$ is derived from a direct utility function of the form $\alpha\log c_N+(1-\alpha)\log c_T$.

In this case,

$$c_{a,t}=(1-\gamma\delta)w_{a,t},$$

and per capita consumption is

$$c_t=(1-\gamma\delta)(1-\gamma)\sum_{a=0}^{\infty}\gamma^a c_{a,t}$$

$$=(1-\gamma\delta)(1-\gamma)\sum_{a=0}^{\infty}\gamma^a w_{a,t}.$$

Using (18), we obtain

$$(19)\quad c_t=(1-\gamma\delta)\left\{\sum_{\tau=0}^{\infty}(\gamma/R)^{\tau}\times\left[y(p_{t+\tau})-\theta_{t+\tau}\right]-Rb_{t-1}\right\},\quad t\geq 0$$

and by aggregating (16) over all age groups we obtain

$$(20)\quad b_t=Rb_{t-1}+c_t-\left[y(p_t)-\theta_t\right],\quad t\geq 0$$

with the initial condition $b_{-1}=\bar{b}_{-1}$.

Finally, we have a market-clearing equation for nontraded goods. Assume that there exist fixed quantities of traded and nontraded goods; y_T and y_N, respectively. In this case,

$$(21)\quad y(p_t)=y_T+p_t y_N,$$

and the equilibrium condition in the market for nontraded goods becomes

$$(22)\quad \alpha c_t=p_t y_N,$$

where the left-hand side represents spending on nontraded goods in terms of traded goods.

The dynamic equations (19)–(20) together with the side conditions (21)–(22) determine a unique equilibrium path that satisfies the solvency constraint:

$$(23)\quad \lim_{t\to\infty}R^{-t}b_t=0.$$

The competitive equilibrium without intervention is obtained by solving (19)–(23) for $g=0$, which implies $\theta_t=0$ for all t. It is shown in the Appendix that this solution is

$$(24)\quad \bar{b}_t=B+(\bar{b}_{-1}-B)\lambda^{t+1}$$

$$(25)\quad \bar{c}_t=C-z(\bar{b}_{-1}-B)\lambda^{t}$$

$$(26)\quad \bar{p}_t=\alpha\bar{c}_t/y_N$$

where

$$(27)\quad B=-y_T(1-\delta R)/\Delta$$

$$(28)\quad C=-y_T(((1-\gamma\delta)(1-\gamma)R)/\gamma\Delta)$$

$$(29)\quad \Delta=\alpha R(1-\gamma\delta)(1-\gamma)/\gamma-(1-\gamma\delta R)(R-\gamma)/\gamma$$

$$(30)\quad \lambda=R\left[\mu-\sqrt{\mu^2-4\delta}\right]/2>0$$

$$(31)\quad \mu=(1-\alpha)(\gamma\delta+1/\gamma)+\alpha(1+\delta)>0$$

$$(32)\quad z=\frac{(1-\gamma\delta)R^2(1-\lambda\gamma/R)}{R[1-\alpha(1-\gamma\delta)]-\lambda\gamma}$$

$$=\frac{R[1-(1-\alpha)(1-\gamma\delta)]-\lambda}{\alpha(1-\alpha)(1-\gamma\delta)}\times(1-\lambda\gamma/R)>0.$$

In order to minimize the number of cases to be investigated, assume that $\bar{b}_{-1}=0$. Define also,

$$(33)\quad \bar{\alpha}=\frac{(1-\gamma\delta R)(R-\gamma)}{R(1-\gamma\delta)(1-\gamma)}.$$

Then we have three cases to consider:

Case 1: $\gamma\delta R>1 \quad\Rightarrow(\delta R>1,\ \alpha>0>\bar{\alpha})$
or $\gamma\delta R<1,\ \delta R>1,\ \alpha>\bar{\alpha} \quad\Rightarrow(1>\bar{\alpha}>0)$

Case 2: $\gamma\delta R<1,\ \delta R>1,\ \alpha<\bar{\alpha}$
$\Rightarrow(1>\bar{\alpha}>0)$

Case 3: $\delta R<1 \quad\Rightarrow(\gamma\delta R<1,\ \bar{\alpha}>1>\alpha)$

Case 1: $B>0$, $C<0$, $\lambda>1$, implying that the country is a creditor in all time periods and its foreign asset holdings are rising over time. Consequently, its aggregate spending is also rising over time, leading to permanent appreciations of the real exchange rate $(1/\bar{p}_t)$. There exists no steady state.

Case 2: $B<0$, $C>0$, $1>\lambda>0$, implying that the properties of the dynamic path are the same as in Case 1, except that the economy converges to a steady state with $\bar{b}_\infty = B<0$, $\bar{c}_\infty = C$ and $\bar{p}_\infty = \alpha C/y_N$.

Case 3: $B>0$, $C>0$, $1>\lambda>0$, implying that the country is a debtor in all time periods, its debt is growing over time, its total spending is declining over time, and its real exchange rate is depreciating in all time periods. The economy approaches a steady state with $\bar{b}_\infty = B>0$, $\bar{c}_\infty = C$ and $\bar{p}_\infty = \alpha C/y_N$. The dynamics of Cases 2 and 3 are depicted in Figures 1 and 2, respectively.

In order to study the comparative dynamics of an exchange rate freeze, we have to compare the solution $(\bar{b}_t, \bar{c}_t, \bar{p}_t)$ with the solution (b_t, c_t, p_t) that obtains for $g>0$. We have not been able to obtain clear-cut comparative-dynamics results from the analytic solution to system (19)–(23) when $\alpha>0$. For this reason we proceed in two stages. First, we present analytically the comparative dynamics for the case $\alpha=0$; that is, the case in which consumers do not value nontraded goods, and then we simulate cases in which $\alpha>0$.

The properties of the solution for $\alpha=0$ is of interest for two reasons. First, continuity implies that the same properties are preserved for $\alpha>0$ but small enough. Hence, it provides insight into comparative dynamics of cases in which the share of spending on nontraded goods is small. Second, our simulations indicate that with large shares of spending on nontraded goods, the effects of an exchange rate freeze are similar to those that we have derived for the case $\alpha=0$.[6]

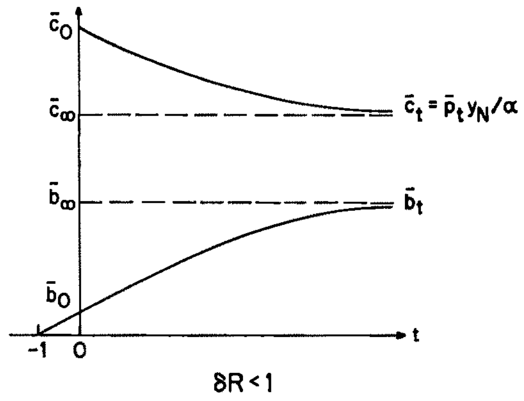


FIGURE 1

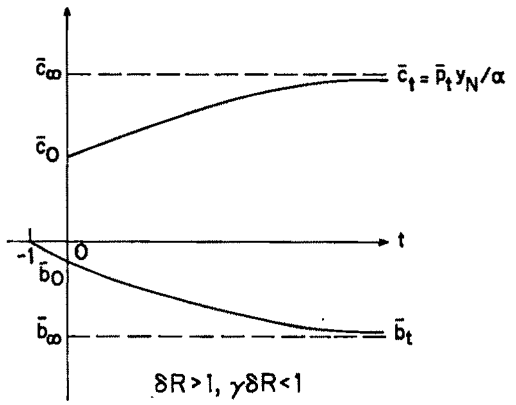


FIGURE 2

Now, solving (19)–(23) for the case $\alpha=0$ and $g>0$ by means of direct iterations of the dynamic equations, we obtain the follow-

[6] The case $\alpha=0$ is simpler than the case $\alpha>0$ because when nontraded goods are not valued by consumers, the dynamic system does not depend on future endogenous variables, and it can therefore be solved by a simple iterative procedure starting at $t=0$. However, when $\alpha>0$, period t wealth depends on the price of nontraded goods in all periods from t to infinity. In this case, the dynamic equations at time t depend on an infinite future sequence of an endogenous variable; i.e., the real exchange rate, requiring a complicated solution procedure. This difficulty is well known in rational expectations models.

ing solution:

$$(34)\quad c_t-\bar{c}_t=$$

$$\begin{cases}(1-\gamma\delta)g\Bigg[1-\dfrac{R-1}{R-\gamma}\gamma^T-\dfrac{R-1}{R-\gamma}(1-\gamma)\\ \quad\times\gamma^T\dfrac{1-(\gamma^2\delta)^{-t}}{1-(\gamma^2\delta)^{-1}}(\gamma^2\delta)^{-1}\Bigg](\gamma\delta R)^t, & 0\le t\le T-1\\ (1-\gamma\delta)g\Bigg[1-\dfrac{R-1}{R-\delta}\gamma^T-\dfrac{R-1}{R-\gamma}(1-\gamma)\\ \quad\times\gamma^T\dfrac{1-(\gamma^2\delta)^{-T+1}}{1-(\gamma^2\delta)^{-1}}(\gamma^2\delta)^{-1}\Bigg](\gamma\delta R)^{T-1}\\ -g(1-\gamma\delta)\dfrac{R-1}{R-\gamma}(1-\gamma)R^T\dfrac{1-(\gamma\delta R)^{t+1-T}}{1-\gamma\delta R}, & t\ge T\end{cases}$$

$$(35)\quad \bar{b}_t-b_t=$$

$$\begin{cases}g\Bigg[(1-\gamma\delta)\dfrac{R-1}{R-\gamma}\gamma^T\dfrac{1-(\gamma^2\delta)^{-t-1}}{1-(\gamma^2\delta)^{-1}}+1\Bigg](\gamma\delta R)^t, & 0\le t\le T-1\\ (\bar{b}_{T-1}-b_{T-1})(\gamma\delta R)^{t+1-T}\\ \quad+g\gamma(1-\delta R)\dfrac{R-1}{R-\gamma}R^{T-1}\dfrac{1-(\gamma\delta R)^{t+1-T}}{1-\gamma\delta R}, & t\ge T\end{cases}$$

Obviously, in this case, $p_t=0$ for all t. It is clear from this solution that spending c_t is larger than $\bar{c}_t$ for $t=0$ and possibly for other small values of t, and that $c_t<\bar{c}_t$ for t large enough. Hence, the exchange rate freeze brings about higher spending levels initially and lower spending levels in the future, as compared to no intervention. Also, when $\delta R<1$ (Case 3), it makes private external debt lower in all time periods. When $\delta R>1$ private debt is lower until some time after T and higher thereafter. Nevertheless, a direct calculation shows that *total* debt $b_t+b_t^G$ is larger than $\bar{b}_t$ in all time periods, which means that public debt increases by more than private debt declines. A comparison of the time profiles of consumption and debt for Case 3 is illustrated in Figure 3.

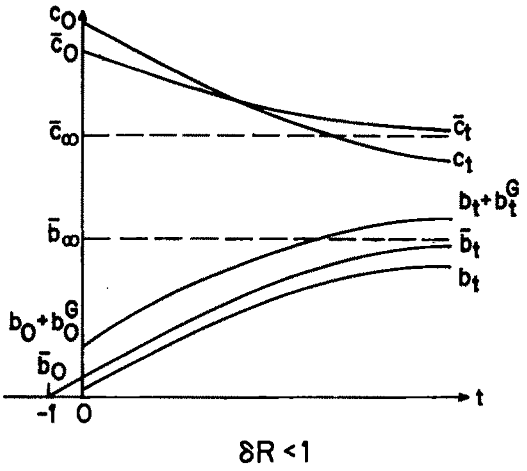


FIGURE 3

It is seen from (34) that the initially higher spending level is larger the later taxes are imposed, and that the eventually lower spending level is smaller the later taxes are imposed. Thus, the longer the delay in the required contractionary policy, the larger are the real effects of the exchange rate freeze. Moreover, the contractionary policy cannot be delayed at will, because given a limit on taxing capacity, say x percent of GNP (possibly 99 percent), taxes which eventually have to equal $g(R-1)R^{T-1}$ cannot exceed x percent of y_T.

The economics behind these results are as follows. The capital gain from the unexpected exchange rate freeze is appropriated by the individuals who are alive in period zero. To them the present value of future tax liabilities is smaller than the capital gain, and they respond by raising spending in all periods. All future generations face larger tax liabilities and reduce spending; the later an individual is born, the larger his tax liability in present value terms (except that all those who are born after $T-1$ have the same tax liability). Over time, the population share of individuals who were alive at $t=0$ declines and the share of those with heavier tax liabilities increases. Therefore, aggregate spending is initially larger and it becomes smaller far enough in the future.

TABLE 2

	b_t		$p_t = .7c_t$		$b_t + b_t^G$
t	$g = 0$	$g = \frac{.1}{(R-1)R^{T-1}}$	$g = 0$	$g = \frac{.1}{(R-1)^{T-1}}$	$g = \frac{.1}{(R-1)^{T-1}}$
0	3.237	2.687	9.887	10.196	3.370
1	4.957	4.413	5.590	5.734	5.164
2	5.871	5.295	3.308	3.361	6.121
3	6.356	5.723	2.096	2.098	6.632
4	6.613	5.905	1.452	1.423	6.905
5	6.750	6.050	1.110	1.060	7.050
6	6.823	6.128	.928	.868	7.128
7	6.862	6.168	.831	.766	7.168
8	6.882	6.190	.780	.711	7.190
9	6.893	6.202	.753	.683	7.202
10	6.899	6.208	.738	.667	7.208
∞	6.905	6.215	.722	.650	7.215

Note: $y_N = y_T = 1$, $b_{-1} = b_{-1}^G = 0$, and $R = 1.1$, $\gamma = .9$, $\delta = .5$, $\alpha = .7$, $T = 5$. Case 3: $\delta R < 1$.

TABLE 3

	b_t		$p_t = .7c_t$		$b_t + b_t^G$
t	$g = 0$	$g = \frac{.1}{(R-1)R^{T-1}}$	$g = 0$	$g = \frac{.1}{(R-1)^{T-1}}$	$g = \frac{.1}{(R-1)R^{T-1}}$
0	−.733	−.739	.622	.623	−.733
1	−1.996	−2.007	1.099	1.099	−1.995
2	−4.169	−4.192	1.919	1.099	−4.167
3	−7.911	−7.957	3.330	3.331	−7.907
4	−14.353	−14.446	5.760	5.761	−14.346
5	−25.445	−25.530	9.945	9.942	−25.430
6	−44.540	−44.614	17.148	17.141	−44.514
7	−77.416	−77.470	29.550	29.536	−77.370
8	−134.018	−134.036	50.903	50.875	−133.936
9	−231.465	−399.091	87.664	87.613	−231.324
10	−399.236	−399.091	150.953	150.864	−398.991
∞	$-\infty$	$-\infty$	$+\infty$	$+\infty$	$-\infty$

Note: $y_N = y_T = 1$, $b_{-1} = b_{-1}^G = 0$, and $R = 2$, $\gamma = .9$, $\delta = .5$, $\alpha = .7$, $T = 5$. Case 1: $\gamma\delta R > 1$.

In order to obtain an idea about the economy's response to an exchange rate freeze when consumers value nontraded goods, we have simulated the system (19)–(23) for various parameter values in order to cover all three cases. The qualitative results of the simulations were always the same; aggregate spending increases on impact and stays higher for some time, it falls below the no-intervention level at some point in time and remains lower thereafter. Naturally, this implies a real depreciation on impact, a lower real exchange rate for some time, and a higher one eventually. In addition, private debt declines on impact while public debt increases, resulting in an increase in aggregate foreign debt (or decline in assets). Private debt remains lower for some time, but may eventually become larger than the no-intervention level of debt. However, aggregate foreign debt is higher in all time periods. These results are in line with the observed response of private consumption, debt, and the real exchange rate to the disinflation attempts in Israel, Chile, and Argentina, that were discussed in the introduction.

Tables 2 and 3 show two simulations that substantiate our claim. The numbers in Table 2 show the same features as the curves in Figure 3. Table 3 tells a similar story for a much higher interest rate. The higher interest rate turns the economy into a foreign asset holder and causes consumption to rise over time instead of declining. However, the comparative dynamics of the exchange rate freeze do not change, except for the fact that private foreign debt becomes higher (foreign asset holdings decline) from period 9 onward, which is larger than $T=5$. This also happens when consumers do not value nontraded goods.

IV. Concluding Comments

Our analysis suggests that it is extremely difficult to disinflate by means of exchange rate management without affecting consumption, debt, and the real exchange rate. When an exchange rate freeze (or a slowdown of devaluation) is imposed without a reduction of private disposable income, the resulting loss of reserves worsens the economy's net asset position vis-à-vis the rest of the world. It also brings about higher current consumption, which is paid for with lower consumption of future generations, and a real exchange rate appreciation. These predictions are consistent with the episodes that were described in the introduction, despite the fact that in those episodes exchange rate management ended with large devaluations, which seems different from the accompanying policies that we have specified. Observe, however, that an expected devaluation acts in our model as a tax on money balances. This is seen from equation (1a) by remembering that $e_{t-1}y(p_{t-1})=m_{t-1}$, so that if θ_t is the effective real tax rate, it is

$$\theta_t = y(p_t)-(m_{t-1}-\tilde{\theta}_t)/e_t.$$

Hence, a higher exchange rate in period t increases the effective tax rate by depleting the real value of money balances; the explicit real tax rate is $\tilde{\theta}_t/e_t$.

It is clear from this discussion—and it can be shown in detail—that if, say, in period T the exchange rate freeze is ended with a maxidevaluation, then the implicit tax on money balances per se will reduce the size of the public debt (will increase reserves). In this case the structure of effective taxes will not be as we have specified, but it is possible to describe every feasible combination of the postfreeze exchange rate and tax policy by a suitable sequence of effective tax rates. The crucial element in our specification is that effective tax payments are postponed when the exchange rate freeze is introduced, and this feature seems to fit well the episodes of our concern.

The predictions of our model are also consistent with the effects of Israel's stabilization program that began in July 1985. The program started with a maxidevaluation of about 19 percent, a reduction of the government's domestic deficit from 11.5 percent of GNP to 4.8 percent of GNP,[7] an exchange rate freeze, and wage and price controls. The reduction in the budget deficit has been achieved mainly by means of an increase in net taxes (taxes minus transfers), from 11.8 percent of GNP in 1984 to 20.4 percent of GNP in 1985 (see Eitan Berglas, 1986). According to our interpretation, the initial devaluation together with the net tax increase brought about a reduction in private consumption, an increase in the real exchange rate (real devaluation), and an improvement in the private sector's trade balance, as is evident from Table 4. This is significantly different from the results of the 1982 episode (see Table 1), in which the exchange rate management was not supported by an absorption policy.

Of course, we do not claim that our model provides a complete explanation of these episodes, nor do we claim that it is the only explanation. We do, however, believe that our analysis highlights certain intertemporal aspects of exchange rate policies that are important in reality, and in particular that

[7] See Bank of Israel, *Annual Report* (1985).

TABLE 4—ISRAEL

			1985		
	1983	1984	I	II	1985
Index of Real Exchange Rate[a] (annual percent change)	−8.2	4.9	5.0	8.6	6.8
Index of Private Consumption[b] (annual percent change)	7.8	−6.8	−0.7	−1.7	−0.4
Trade Balance[c] (billions of U.S. dollars)	−2.6	−1.5	−0.14	−0.13	−0.26

[a] Implicit Price Deflator for Imports, relative to the Implicit Price Deflator of Domestic Absorption: Source: Bank of Israel *Annual Report*, 1985, Table G-1 (Hebrew).
[b] Source: Bank of Israel *Annual Report*, 1985, Table A-1 (Hebrew).
[c] Current Account Balance, Minus Interest Payments, Minus Defense Imports. Source: Bank of Israel *Annual Report*, 1985, Table G-1 (Hebrew).

exchange rate management contains important tax-equivalent elements. These taxation effects are distributed over time, depending on the exchange rate policy, and they can be offset by regular taxes. Hence, there exist close links between exchange rate management and fiscal policies.

APPENDIX

Here we derive the solution of (19)–(23) for $g=0$ (i.e., equations (24)–(26)), and the general solution for $g>0$. Then we explain the calculation method used in Tables 2 and 3.

By substituting (21) and (22) into (19)–(20) we obtain

$$\text{(A1)}\quad p_t y_N = \alpha(1-\gamma\delta) \times\left[\frac{R}{R-\gamma}y_T + \sum_{\tau=0}^{\infty}\left(\frac{\gamma}{R}\right)^{\tau} \times(p_{t+\tau}y_N-\theta_{t+\tau}) - Rb_{t-1}\right]$$

$$\text{(A2)}\quad b_t = Rb_{t-1} + \frac{1}{\alpha}p_t y_N - y_T - p_t y_N + \theta_t.$$

Now define

$$\text{(A3)}\quad P_t = \sum_{\tau=0}^{\infty}\left(\frac{\gamma}{R}\right)^{\tau} p_{t+1+\tau},\qquad t=-1,0,1,2,\ldots,$$

which implies

$$\text{(A4)}\quad p_t = P_{t-1} - \frac{\gamma}{R}P_t,\quad t=0,1,2,\ldots.$$

Substituting (A4) into (A1) and (A2), we obtain the following system of difference equations:

$$\text{(A5)}\quad \begin{bmatrix} P_t \\ b_t \end{bmatrix} = A\begin{bmatrix} P_{t-1} \\ b_{t-1} \end{bmatrix} + y_T a + D_t;\qquad t=0,1,2,\ldots,$$

where

$$\text{(A6)}\quad A = \begin{bmatrix} [1-\alpha(1-\gamma\delta)]R/\gamma & \alpha(1-\gamma\delta)\dfrac{R^2}{\gamma y_N} \\ (1-\alpha)(1-\gamma\delta)y_N & R[1-(1-\alpha)(1-\gamma\delta)] \end{bmatrix}$$

$$\text{(A7)}\quad a = \begin{bmatrix} -\dfrac{\alpha(1-\gamma\delta)R^2}{\gamma y_N(R-\gamma)} \\ -1+(1-\alpha)(1-\gamma\delta)\dfrac{R}{R-\gamma} \end{bmatrix}$$

$$\text{(A8)}\quad D_t = \begin{bmatrix} \dfrac{\alpha R(1-\gamma\delta)}{\gamma y_N}\displaystyle\sum_{\tau=0}^{\infty}\left(\frac{\gamma}{R}\right)^{\tau}\theta_{t+\tau} \\ \theta_t - (1-\alpha)(1-\gamma\delta) \times\displaystyle\sum_{\tau=0}^{\infty}\left(\frac{\gamma}{R}\right)^{\tau}\theta_{t+\tau} \end{bmatrix}$$

where θ_t is defined in (17). We also have an initial and a terminal condition. The initial condition is on b_{-1} and the terminal condition is (23) in the text; that is,

$$b_{-1} = \bar{b}_{-1} \tag{A9}$$

$$\lim_{t \to \infty} R^{-t} b_t = 0. \tag{A10}$$

Now, the no-intervention solution is obtained by choosing $g = 0$, which implies $\theta_t = 0$ and $D_t = 0$ for all t. In this case, (A5) is a homogeneous system whose general solution is

$$\begin{bmatrix} \bar{P}_t \\ \bar{b}_t \end{bmatrix} = \begin{bmatrix} P \\ B \end{bmatrix} + A^{t+1} \begin{bmatrix} \bar{P}_{-1} \\ \bar{b}_{-1} \end{bmatrix} \tag{A11}$$

where

$$\begin{aligned} \begin{bmatrix} P \\ B \end{bmatrix} &= y_T (I - A)^{-1} a \\ &= -\frac{y_T}{\Delta} \begin{bmatrix} \dfrac{\alpha(1-\gamma\delta)(1-\gamma)R^2}{\gamma y_N (R-\gamma)} \\ 1 - \delta R \end{bmatrix} \end{aligned} \tag{A12}$$

$$\begin{aligned} \Delta = {} & \alpha R (1-\gamma\delta)(1-\gamma)/\gamma \\ & - (1 - \gamma\delta R)(R - \gamma)/\gamma. \end{aligned} \tag{A13}$$

However,

$$A^t = V \Lambda^t V^{-1}, \tag{A14}$$

where Λ is a diagonal matrix with the eigenvalues of A on the diagonal and V is a matrix of the corresponding eigenvectors. We normalize so that

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & v_2 \\ 1 & 1 \end{bmatrix}$$

where

$$\lambda_i = \frac{R}{2} \left[\mu + (-1)^i \sqrt{\mu^2 - 4\delta} \right], \quad i = 1,2 \tag{A15}$$

$$\mu = (1 - \alpha)\left(\frac{1}{\gamma} + \gamma\delta \right) + \alpha(1 + \delta). \tag{A16}$$

Then,

$$\begin{aligned} v_i &= \frac{\alpha(1-\gamma\delta)R^2}{y_N \{ \gamma\lambda_i - R[1 - \alpha(1 - \gamma\delta)] \}} \\ &= \frac{\lambda_i - R[1 - (1-\alpha)(1-\gamma\delta)]}{(1-\alpha)(1-\gamma\delta) y_N}, \quad i = 1,2. \end{aligned} \tag{A17}$$

Substitution of (A12) and (A14) into (A11) gives us

$$\begin{aligned} \bar{P}_t = {} & P + (1/(v_2 - v_1)) \\ & \times \{ [v_2(\bar{P}_{-1} - P) - v_1 v_2 (\bar{b}_{-1} - B)] \lambda_2^{t+1} \\ & - [v_1(\bar{P}_{-1} - P) - v_1 v_2 (\bar{b}_{-1} - B)] \lambda_1^{t+1} \} \end{aligned} \tag{A18a}$$

$$\begin{aligned} \bar{b}_t = {} & B + (1/(v_2 - v_1)) \\ & \times \{ [\bar{P}_{-1} - P - v_1(\bar{b}_{-1} - B)] \lambda_2^{t+1} \\ & - [\bar{P}_{-1} - P - v_2(\bar{b}_{-1} - B)] \lambda_1^{t+1} \}. \end{aligned} \tag{A18b}$$

Now, since $\mu > 1 + \delta$ and λ_2 is increasing in μ, then

$$\lambda_2 > \frac{R}{2} \left[(1+\delta) + \sqrt{(1+\delta)^2 - 4\delta} \right] = R.$$

On the other hand, λ_1 is declining in μ. Therefore

$$\lambda_1 < \frac{R}{2} \left[(1+\delta) - \sqrt{(1+\delta)^2 - 4\delta} \right] = \delta R < R.$$

Hence, since $v_1 < 0$ and $v_2 > 0$ the terminal condition (A10) is satisfied by (A18b) if and only if:

$$\bar{P}_{-1} = P + v_1(\bar{b}_{-1} - B). \tag{A19}$$

This determines the initial value of P_{-1}. Substituting (A19) into (A18) yields

$$\bar{P}_t = P + v_1(\bar{b}_{-1} - B)\lambda^{t+1}, \tag{A20a}$$

$$\bar{b}_t = B + (\bar{b}_{-1} - B)\lambda^{t+1}, \tag{A20b}$$

where $\lambda = \lambda_1$. Equation (A20b) is presented in the text as equation (24).

In order to obtain a solution to $(\bar{p}_t, \bar{c}_t)$, we first use (A4) and (A20a) to calculate

$$\text{(A21)}\quad \bar{p}_t = \left(1-\frac{\gamma}{R}\right)P + v_1(\bar{b}_{-1} - B) \times\left(1-\frac{\lambda\gamma}{R}\right)\lambda^t.$$

Then, substituting (A21) into (22), we obtain

$$\bar{c}_t = \frac{y_N}{\alpha}\left(1-\frac{\gamma}{R}\right)P + \frac{y_N v_1}{\alpha}\left(1-\frac{\lambda\gamma}{R}\right)(\bar{b}_{-1} - B)\lambda^t.$$

Hence

$$\text{(A22)}\qquad \bar{c}_t = C - z(\bar{b}_{-1} - B)\lambda^t,$$

where $C = y_N(1-\gamma/R)P/\alpha$ and $z = -y_N v_1(1-\lambda\gamma/R)/\alpha$. This is equation (25).

In order to solve the system for the case $g > 0$, observe that

$$\text{(A23)}\quad D_t = D = -(R-1)R^{T-1}ga, \text{ for } t \geq T$$

Hence, for $t \geq T$ equation (A5) describes a homogeneous system for which we can apply the same procedure that was used to calculate $(\bar{b}_t, \bar{p}_t)$. This yields the counterparts of (A20):

$$\text{(A24a)}\quad P_t = P_g + v_1(b_T - B_g)\lambda^{t-T}, \quad t \geq T,$$

$$\text{(A24b)}\quad b_t = B_g + (b_T - B_g)\lambda^{t-T}, \quad t \geq T,$$

where

$$\text{(A25)}\quad \begin{bmatrix} P_g \\ B_g \end{bmatrix} = -\frac{1}{\Delta}\left[y_T - (R-1)R^{T-1}g\right] \times \begin{bmatrix} \dfrac{\alpha(1-\gamma\delta)(1-\gamma)R^2}{\gamma y_N(R-\gamma)} \\ 1-\delta R \end{bmatrix}$$

Using forward iterations of (A5) in order to calculate (P_T, b_T) as functions of (P_{-1}, g), and (A24a) evaluated at $t = T$, we obtain a system of three linear equations which enable us to solve (P_T, b_T, P_{-1}) as functions of g. Having calculated P_{-1} we use (A5) to calculate (P_t, b_t) for $0 \leq t \leq T$ and (A24) to calculate (P_t, b_t) for $t \geq T$. Finally, having calculated the sequence $\{P_t, b_t\}_{t=-1}^{\infty}$ we calculate $\{p_t, c_t\}_{t=0}^{\infty}$ using (A4) and (22). Tables 2 and 3 were computed by means of this procedure.

REFERENCES

Aschauer, David and Greenwood, Jeremy, "A Further Exploration in the Theory of Exchange Rate Regimes," *Journal of Political Economy*, October 1983, *91*, 868–75.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

Berglas, Eitan, "Taxes and Transfers: 1975–1985," Discussion Paper No. 12–86, Sapir Center for Development, Tel Aviv University, September 1986.

Blanchard, Olivier, J., "Debt, Deficits and Finite Horizons," *Journal of Political Economy*, April 1985, *93*, 223–47.

Dornbusch, Rudiger, "External Debt, Budget Deficits, and Disequilibrium Exchange Rates," in G. W. Smith and J. T. Cuddington, eds., *International Debt and the Developing Countries*, Washington: World Bank, 1985.

Drazen, Allan and Helpman, Elhanan, "Exchange Rate Management and Monetary and Fiscal Policy," mimeo., June 1986.

Feenstra, Robert C., "Functional Equivalence between Liquidity Costs and the Utility of Money," *Journal of Monetary Economics*, May 1986, *17*, 271–91.

Frenkel, Jacob A. and Razin, Assaf, "Budget Deficits and Rates of Interest in the World Economy," *Journal of Political Economy*, June 1986, *94*, 564–94.

Harberger, Arnold, C., "The Chilean Economy in the 1970s: Crisis, Stabilization, Liberalization, Reform," *Carnegie-Rochester Series on Public Policy*: *Economic Policy in a World of Change*, Autumn 1982, *17*, 115–52.

Helpman, Elhanan, "An Exploration in the

Theory of Exchange Rate Regimes," *Journal of Political Economy*, October 1981, *89*, 865–90.

______ **and Razin, Assaf,** "Towards a Consistent Comparison of Alternative Exchange Rate Regimes," *Canadian Journal of Economics*, August 1979, *12*, 394–409.

______ **and** ______, "The Role of Saving and Investment in Exchange Rate Determination under Alternative Monetary Mechanisms," *Journal of Monetary Economics*, May 1984, *13*, 307–25.

Lucas, Robert, E., Jr., "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, *10*, 335–60.

Obstfeld, Maurice, "The Capital Inflows Problem Revisited: A Stylized Model of Southern Cone Disinflation," *Review of Economic Studies*, October 1985, *52*, 605–26.

______, "Speculative Attack and the External Constraint in a Maximizing Model of the Balance of Payments," *Canadian Journal of Economics*, February 1986, *19*, 1–22.

Polak, J. J., "Monetary Analysis of Income Formation and Payments Problems," *IMF Staff Papers*, November 1957, *6*, 1–50.

van Wijnbergen, Sweder, "Fiscal Deficits, Exchange Rate Crises and Inflation: On the Inflationary Consequences of Anti-Inflationary Exchange Rate Policies," Working Paper No. 2–86, Foerder Institute for Economic Research, Tel Aviv University, January 1986.

Bank of Israel, *Annual Report*, 1985.

Israel Central Bureau of Statistics, *National Accounts 1972–1975*, Appendix to *Israel Statistical Monthly*, No. 5, 1986.

International Monetary Fund, *International Financial Statistics*, December 1984, *37*.

The Global Correspondence Principle: A Generalization

By JAGDISH N. BHAGWATI, RICHARD A. BRECHER, AND TATSUO HATTA*

This paper generalizes the Global Correspondence Principle by extending, in two major ways, Paul Samuelson's 1971 analysis of the exchange rate response to an international purchasing-power transfer. We analyze the price effect of a shift in any parameter, not necessarily a transfer. We then explore the resulting adjustments in any nonprice variable such as welfare. As our analysis shows, the direction of these adjustments depends neither on whether they are small or large nor on whether equilibrium is locally stable or unstable.

The celebrated Correspondence Principle of Paul Samuelson (1947) highlighted the importance of local stability for deriving fruitful theorems in comparative statics based on "small" changes. Thus, theorists typically qualify their comparative-static conclusions with the proviso that equilibrium is locally stable.

As Samuelson (1971) subsequently pointed out, however, the comparative-static effects of a parametric shift upon relative prices do not depend qualitatively on whether the initial equilibrium is locally Walras stable, provided that Walrasian tâtonnement is invoked and "large" adjustments are taken properly into account. Thus, using Samuelson's recent (1983) terminology, we have the Global Correspondence Principle.

Samuelson's 1971 paper dealt explicitly with the case of an international purchasing-power transfer in a two-country, two-good model. He examined the effect of a transfer upon the exchange rate, that is, the price variable whose disequilibrium behavior is directly specified by the adjustment mechanism postulated in the model.[1]

First, we shall analyze the global response of price to a shift in *any* parameter (not necessarily a transfer payment) within the two-good, general equilibrium model. Then, more remarkably, we shall also establish that the corresponding responses of *nonprice* variables, such as welfare, are qualitatively independent of both the magnitude of adjustment and the question of local stability. This result extends the Global Correspondence Principle to cover variables other than price. The extension enables us to prove that, in a two-agent, two-good world, a transfer of income always improves the recipient's welfare—regardless of whether the initial equilibrium is stable, and irrespective of the magnitude of the transfer. In this way we generalize Samuelson's (1947) well-known proposition that the recipient's welfare improves provided that the initial equilibrium is stable. To take another example, our analysis can also be utilized to show that local stability can be dropped from Bhagwati's (1958) well-known set of conditions for immiserizing growth.

In Section I, we set up the model and analyze the global effect of a parametric shift on the relative price. In Section II, the shift's

*Bhagwati: Department of Economics, Columbia University, New York, NY 10027; Brecher: Department of Economics, Carleton University, Ottawa, ON K1S 5B6, Canada; Hatta: The Institute of Social and Economic Research, Osaka University, 6-1, Mihogaoka, Ibaraki, Osaka 567, Japan. Thanks are due to Kenneth Arrow, Bela Balassa, Franklin Fisher, Suezo Ishizawa, Murray Kemp, Takashi Negishi, Peter Newman, Zvi Safra, Paul Samuelson, and T.N. Srinivasan for useful comments on this as well as an earlier draft, whose main thrust has been substantially changed and broadened in the present paper as noted below. Partial financial support for Bhagwati's research was provided by the National Science Foundation.

[1] We may mention that the original draft of our present paper also happened to develop a global analysis of price changes in relation to the transfer problem. Samuelson was kind enough to draw our attention to the fact that his 1971 paper had already anticipated our findings. The present draft is thus focused on two other aspects of global analysis, as explained in the text.

effect on nonprice variables will be studied and our main results proved. Section III applies our results of Section II specifically to the example of a transfer payment. Concluding remarks are offered in Section IV.

I. Price Effects

We adopt the standard competitive model of an economy in which two goods, X and Y, are produced and consumed. Let the economy's aggregate excess demand function for good X be $x(p,\theta)$, where p is the relative price of this good, and θ is a shift parameter. This generic model can be interpreted in many different ways. For example, p and x may be domestic price and excess demand for X in a single-country economy, with θ representing a sales tax, an internal transfer between agents of the country, or the endowment of a factor. Alternatively, in the case of more than one country, p and x may be the international terms of trade and world excess demand, while θ could be an import tariff, an international transfer or domestic productivity.

The condition for market equilibrium is

$$(1) \qquad x(p,\theta)=0.$$

To guarantee existence of a positive equilibrium price, we make the following assumption:

ASSUMPTION 1: *The function $x(p,\theta)$ is continuous with respect to each of its arguments, and for each value of θ in the relevant domain there exist a $\bar{\bar{p}}$ and a $\bar{p}$ such that $x(p,\theta)<0$ for all $p \geq \bar{\bar{p}}$ and $x(p,\theta)>0$ for all $p \leq \bar{p}$.*

To cover situations of disequilibrium, we postulate a dynamic adjustment process of Walrasian tâtonnement, characterized by the following assumption:

ASSUMPTION 2: *When market-clearing condition* (1) *is not satisfied, p continuously increases or decreases as $x(p,\theta) \gtrless 0$, respectively, ceasing to change when $x(p,\theta)=0$.*

We can now show that the direction of change in the price ratio depends only on the sign of the aggregate excess demand created by the parametric shift at the initial equilibrium price. This result is stated as the following proposition:

PROPOSITION 1: *In the model characterized by equation* (1) *and Assumptions 1 and 2, suppose that $x(p^0,\theta^0)=0$ for some pair (p^0,θ^0) of p and θ. Let θ be shifted from θ^0 to θ' where $x(p^0,\theta')\neq 0$. Then the price will reach a new equilibrium value, denoted p', such that*

$$(2) \qquad (p'-p^0)x(p^0,\theta')>0.$$

PROOF:

First suppose that

$$(3) \qquad x(p^0,\theta')<0.$$

Then, from Assumption 2, p' must be the maximum p satisfying both

$$(4) \qquad x(p,\theta')=0;$$

$$(5) \qquad p<p^0.$$

Since Assumption 1 ensures that $x(\bar{p},\theta')>0$ for a $\bar{p}<p^0$, the Intermediate Value Theorem and inequality (3) together imply that there is at least one value of p satisfying both conditions (4) and (5). Denoting the highest such value p', and invoking Assumption 2, we find that p' is the new equilibrium price. Thus, inequalities (3) and (5) yield (2). Similar reasoning establishes this proposition in the alternative case where $x(p^0,\theta')>0$ instead of inequality (3).

Proposition 1 states that the price increases if and only if the parametric shift creates a positive excess demand at the initial equilibrium price. Moreover, this result holds for large as well as small parametric changes of any kind, and regardless of whether local stability obtains, in our two-good model with Walrasian tâtonnement. Correspondingly, the conditions that have been derived in the comparative-static theoretical literature for determining price change from the sign of $x(p^0, \theta')$ following specific parametric shifts, and that have been considered valid only for small changes from a locally Walras-stable equilibrium, are equally valid globally. That is, these conditions hold for changes of any size, with or without local stability. For example, the well-known transfer problem criterion for price adjustments—according to which the donor's terms of trade will improve or worsen as the sum of the marginal propensities to consume importables of the two countries is respectively greater or less than unity[2]—is immediately seen to hold globally, and not just for small changes from a locally stable equilibrium as conventionally established.

Figure 1 illustrates why Proposition 1 holds even when the initial equilibrium is unstable. The solid and dashed curves respectively represent the aggregate excess demand function for good X after and before the parametric shift. Since the dashed curve is drawn upward sloping at p^0, the initial equilibrium is unstable. (To avoid cluttering the diagram, most of this curve is not shown.) At p^0, the shift in θ then creates an excess supply of good X in the case depicted. Thus, the price of this good falls by Walrasian tâtonnement to p', the new equilibrium.[3]

By contrast, the traditional technique of differential calculus evidently cannot be used to infer the shift of an unstable equilibrium, since a small change in θ at such an equilibrium will lead to a large adjustment in p. The calculus technique leads to the following well-known formula derived from equation (1):

$$dp/d\theta = -x_\theta/x_p, \tag{6}$$

where subscripts denote partial derivatives. At an unstable equilibrium, we have $x_p > 0$,

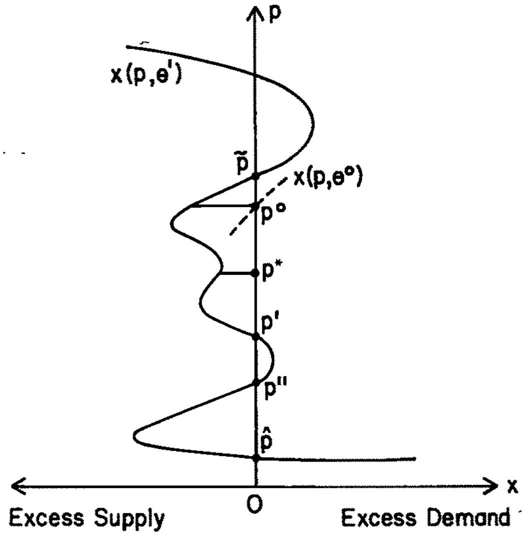


FIGURE 1

in which case $dp/d\theta > 0$ when $x_\theta < 0$. Thus, in Figure 1, we can interpret formula (6) as simply telling us that a price rise from p^0 to $\tilde{p}$ would be necessary to eliminate the excess supply created (at the initial price) by a parametric shift from θ^0 to θ'. It would, however, be wrong to infer, as some statements in the literature by many can suggest, that $\tilde{p}$ will in fact be the new equilibrium in this case. Given the Walrasian price-adjustment mechanism, p will actually fall in response to the excess supply. Hence $\tilde{p}$ will not be the new equilibrium price; rather, the price will adjust globally to p' in this case.

To put the matter somewhat differently, contrary to customary interpretations, it is *sufficient* to look at the sign on the *numerator* alone in formula (6) to determine the direction of change in the terms of trade, in the event of a small parametric shift. For x_θ, the numerator, is nothing but the excess demand at initial terms of trade when θ shifts; that is, it corresponds to $x(p^0, \theta')$ in formula (2). As Proposition 1 shows, moreover, the sign of this excess demand exclusively determines the direction of change in the terms of trade, that is, the sign of $(p' - p^0)$ in formula (2). Thus, the sign of the denominator in formula (6), that is, of x_p, or equivalently the stability term, is *irrelevant*

[2] Of course, this condition on marginal propensities determines the sign of excess demand created at the initial equilibrium price (i.e., the sign of $x(p^0, \theta')$) when the specific parametric shift is a transfer.

[3] If the solid curve were slightly redrawn to touch (rather than cross) the vertical axis at p'—in which case the new equilibrium would no longer be fully stable—the essence of our above argument would be unaffected.

to the direction of price change. The conventional criteria for terms-of-trade adjustment under different parametric shifts, which relate to the sign of the numerator in formula (6) and have been stated as valid subject to Walras-stability restriction to sign the denominator in formula (6), are therefore valid quite generally.

II. Welfare and Other Nonprice Effects

Proposition 1, like Samuelson's Global Correspondence Principle, characterizes the global effect of a parametric shift upon the price variable. An intriguing question is: can the Global Correspondence Principle also be extended to the effects on nonprice variables? For example, the criteria for determining the effects of transfers or growth on a country's welfare have been established, invoking again Walras stability. Can these results also be shown to be independent of such a stability restriction? If so, the demonstration would be truly remarkable, since the dynamic adjustment behavior is specified on prices, not on the nonprice variables.

In this section, we indeed demonstrate that the criteria established for changes in nonprice variables in response to a small parametric shift, using differential calculus methods and considered valid in the presence of Walras stability, are valid in the global context under certain reasonable conditions.

To carry out this demonstration, we resort to an analogue of the technique utilized in the price-change problem above, where the excess demand induced by the parametric shift at initial prices (i.e., $x(p^0,\theta')$) played the key role. Since the objective now is to analyze nonprice change instead, we reassign this role to excess demand at the initial value of the nonprice variable. To maintain the value of the nonprice variable at the initial level, of course, we shall assume that price adjusts as required (to a level to be denoted as p^* below). Thus, for example, if welfare is the nonprice variable and domestic growth is the shift parameter, the focus now would be on the excess demand where the price has been adjusted (to p^*) to leave welfare unchanged after growth.

The criteria derived by examining the factors governing the sign of this excess demand at constant value of the nonprice variable correspond then to the ones stated as the conventional comparative-static results, using differential calculus methods.[4] Although these traditional results are stated as valid provided that Walras stability obtains, our analysis below will demonstrate that the results so derived are valid globally and independently of local stability, under specific restrictions.

Our demonstration proceeds by first deriving the conventional differential calculus results for nonprice-variable change in a general form, for any shift parameter, equivalently to the price-change formula (6). We then examine the precise manner in which these results generalize in the global context, that is, for large changes and regardless of Walras stability.

A. *Local Effects Reconsidered*

First, we introduce the necessary terminology. Let u denote the nonprice variable that we examine, and suppose that it is functionally dependent on both p and θ, as follows:

$$u=v(p,\theta). \tag{7}$$

We call the direct impact of θ upon u the "primary impact" (given by $v(p^0,\theta')-v(p^0,\theta^0)$) and the indirect impact of θ upon u through p the "secondary impact" ($v(p',\theta')-v(p^0,\theta')$). When primary and secondary impacts are opposite in direction and the latter outweighs the former, we shall say that the overall effect on u is "paradoxical." Otherwise, the overall effect will be called "normal."

[4] It should be stated that the comparative-static results on welfare have typically been derived directly, rather than by using the technique of investigating excess demand at constant value of the nonprice variable—a technique that we use to advantage here in generalizing the Global Correspondence Principle to the nonprice domain. Bhagwati's (1958) analysis of immiserizing growth (and, more recently, several analyses of international transfers) did, however, use the technique that we further develop here.

For example, we may interpret u and θ as the welfare level and the transfer receipt, respectively, of one country in the standard model of international trade. Then the primary impact is the direct effect of the transfer upon the recipient's welfare at constant terms of trade, and the secondary impact is the transfer-induced terms-of-trade effect. Since the primary effect increases u in this case, the paradox implies a terms-of-trade deterioration so great as to make the overall effect on u negative. Other examples of paradoxes defined in this way are Bhagwati's immiserizing growth, or a welfare loss from abolishing a tariff.

Now, define the function $\bar{x}$ by

$$(8)\qquad \bar{x}(p,u) \equiv x[p,c(p,u)],$$

where $\theta = c(p,u)$ is the inverse function of equation (7) with respect to θ.[5]

Then, in view of equation (1) and identity (8), the equilibrium values of p and u must satisfy $\bar{x}(p,u)=0$. Thus, we obtain $du/dp = -\bar{x}_p/\bar{x}_u$. This result and equation (6) then yield $du/d\theta = (du/dp)dp/d\theta = x_\theta \bar{x}_p/\bar{x}_u x_p$. Hence, noting that identity (8) implies that $\bar{x}_u = x_\theta c_u$, we have $(du/d\theta)c_u = \bar{x}_p/x_p$. Since the sign of c_u indicates the direction of the primary impact, a paradox implies that $(du/d\theta)c_u < 0$. Now, $x_p < 0$ is the Walras-stability assumption. Therefore, we get the central result that, for small changes in the presence of local stability,

$$(9)\qquad \bar{x}_p > 0$$

is a necessary and sufficient condition for a paradoxical outcome. In other words, $\bar{x}_p \leqq 0$ is a necessary and sufficient condition for a normal outcome, when Walras stability is assumed.

Analyses of specific parametric shifts then relate to $\bar{x}_p$, which is the shift-induced change in excess demand at constant u. Thus, Bhagwati, who analyzed the paradox of immiserizing growth, found that $\bar{x}_p > 0$ could arise from ultra-biased expansion or inelastic demand, despite Walras stability. Moreover, to obtain Samuelson's (1947) result that a transfer between two agents engaged in free trade could not yield welfare paradoxes in the presence of Walras stability, we could show that necessarily $\bar{x}_p < 0$ in this case.

B. *Global Effects*

Can formula (9) be generalized to cover both large (as well as small) parametric changes *and* situations where Walras stability does not obtain? In the general case, we shall show (via Proposition 2) that when (9) is suitably rewritten in equivalent noncalculus terms, it becomes only a necessary (but not sufficient) condition for a paradox to arise (and hence $\bar{x}_p \leqq 0$ becomes correspondingly a sufficient, but not necessary, condition for a normal outcome). Also, if we add certain restrictions (to rule out multiple values of p^* and to ensure monotonicity in the relationship of v to p), the noncalculus version of (9) can indeed be shown (via Lemma 2 and Proposition 3) to become a necessary *and* sufficient condition for a paradox.

We proceed by first obtaining the following lemma:

LEMMA 1: *A necessary condition for the overall effect of θ upon u to be paradoxical is that*

$$(10)\qquad (p' - p^*)(p^* - p^0) > 0,$$

for a p^* satisfying

$$(11)\qquad v(p^*,\theta') = v(p^0,\theta^0).$$

PROOF:

Without loss of generality, let us assume that the primary impact of the parametric shift is positive in the sense that $v(p^0,\theta') > v(p^0,\theta^0)$. When a paradox occurs, we must then have

$$(12)\quad v(p^0,\theta') > v(p^0,\theta^0) > v(p',\theta').$$

[5] When equation (7) can be interpreted as an indirect utility function, $c(p,u)$ is the corresponding expenditure function. When θ is interpreted as the transfer receipt and u the welfare of the recipient agent, then $c(p,u)$ is the expenditure function minus the revenue function of this agent; we called this difference the *overspending function* in our 1983 article.

This result and the Intermediate Value Theorem imply that there must be a p^* between p^0 and p' such that equation (11) is satisfied. Thus, we have either $p^0 < p^* < p'$ or $p' < p^* < p^0$, which implies inequality (10).

Equation (11) indicates that p^* is a price that would leave welfare at the initial level after the shift in θ. Inequality (10) implies that such a p^* must exist between p^0 and p'. In other words, when a paradox occurs as a result of a parametric shift, the price must pass through p^* before it reaches the final equilibrium level.

Intuitive appreciation of this proposition can be strengthened with the help of Figure 1. Inequalities (12) imply that, after the shift in θ, the level of u at p^0 is higher than the initial level, while that at p' is lower. Hence, between p^0 and p', there must be some level of p (denoted p^*) causing equation (11) to hold and making the level of u equal to the initial one.

Lemma 1 yields the following:

PROPOSITION 2: *A necessary condition for the overall effect of θ upon u to be paradoxical is that*

$$(13) \qquad (p' - p^0)x(p^*,\theta') > 0,$$

for a p^* satisfying equation (11).

PROOF:

Take the p^* defined in Lemma 1. Then, from the lemma, p^* is between p^0 and p'. Thus, in view of Assumption 2, the sign of the excess demand $x(p^*,\theta')$ at p^* must be equal to that of $x(p^0,\theta')$ created by the direct impact of the parametric shift. This result and inequality (2) in Proposition 1 immediately yield (13).

Inequality (13) in this proposition—a result that holds evidently for large and small changes regardless of local stability—can now be formally related to $\bar{x}_p$ in (9).

Since $x(p^0,\theta^0) = 0$, and since $(p^* - p^0)$ and $(p' - p^0)$ have the same sign when a paradox occurs (in view of Lemma 1), inequality (13) can be alternatively written as

$$(14) \qquad [x(p^*,\theta') - x(p^0,\theta^0)] \\ /(p^* - p^0) > 0.$$

Recalling identity (8), we can then rewrite condition (14) and hence (13) as

$$(15) \qquad [\bar{x}(p^*,u^0) - \bar{x}(p^0,u^0)] \\ /(p^* - p^0) > 0,$$

where the initial value of u is denoted $u^0[\equiv v(p^0,\theta^0)]$.[6] Condition (9), however, is nothing but the calculus version of (15) and hence of (13) in Proposition 2.[7]

Therefore, it follows from Proposition 2 that the conventional necessary condition for paradoxical nonprice adjustments (derived for small changes and local stability) holds equally for large changes and regardless of stability. This proposition immediately implies that Bhagwati's necessary condition for immiserizing growth holds for changes of any magnitude, with or without stability. A similar generalization can be developed for Samuelson's (1947) theorem regarding the impossibility of a welfare-paradoxical transfer (as in Section III below).

Our generalization of the conventional comparative-static results is, however, not total. While (9) is both a necessary and sufficient condition for a paradox with small changes and local stability, its global counterpart in terms of condition (13) is only necessary but not sufficient.

A full generalization is possible, however, if added restrictions are imposed. These are suggested by noting that there are two reasons why necessary condition (13) is not also

[6] Note that from identity (8) we have $\bar{x}(p^*,u^0) \equiv x[p^*,c(p^*,u^0)] = x(p^*,\theta')$.

[7] Inequality (15) implies that, according to Proposition 2, $\bar{x}$ is (on average) an increasing function with respect to p in the interval between p^0 and p^* when the overall effect of θ upon u is paradoxical. That is, as the price is increased (decreased) from p^0 to p^*, the excess demand has to increase (decrease) if θ is adjusted to keep u constant at u^0.

a sufficient condition for a paradox. First, even when (13) holds, p^* may fail to satisfy inequality (10) in Lemma 1; that is, p^* may not lie between p^0 and p'. Hence, a normal outcome may arise. This situation would occur in Figure 1 if p^* were relocated to lie between p'' and $\hat{p}$, still leaving $x(p^*,\theta')<0$.

Second, even if inequality (13) holds for a p^* satisfying (10), a normal result may arise if $v(p,\theta')$ is not monotonic with respect to p. To see this possibility in Figure 1, note that (by definition) the price movement from p^0 to p^* keeping θ at θ' would just offset the primary impact on u. Thus, if $v(p,\theta')$ is monotonic with respect to p, the further price movement from p^* to p' will continue to change u in the same direction; that is, the secondary will now outweigh the primary impact, necessarily creating a paradox. This outcome need not arise, however, if $v(p,\theta')$ is not monotonic, as may happen in the case when u represents welfare and distortions are present.

Thus, we immediately have the following:

LEMMA 2: *Assume that $v(p,\theta')$ is monotonic with respect to p in the interval between p^0 and p'. Then a necessary and sufficient condition for the overall effect of θ upon u to be paradoxical is that inequality* (10) *hold for a p^* satisfying equation* (11).

This lemma yields the following:

PROPOSITION 3: *Assume that*: (i) *the function $v(p,\theta')$ is monotonic with respect to p, so that p^* is unique*; (ii) *the primary and secondary impacts of the shift from θ^0 to θ' are opposite in direction* (*since otherwise a paradox would clearly be impossible*); *and* (iii) *the shift from θ^0 to θ' is small enough to insure that $x(p,\theta')$ does not change sign more than once between p^0 and p^*. Then, inequality* (13) *is a necessary and sufficient condition for the overall effect of θ upon u to be paradoxical.*

PROOF:

In view of Proposition 2 and Lemma 2, we only have to prove that inequality (13) implies (10) under the assumptions of the present proposition. Now suppose that the function $x(p,\theta')$ changes sign between p^0 and p^* exactly once. Then we have $x(p^0,\theta')x(p^*,\theta')\leqq 0$. This result and inequality (13) yield $(p'-p^0)x(p^0,\theta')\leqq 0$, a contradiction of (2). In view of Proposition 3, assumption (iii), therefore, the function $x(p,\theta')$ cannot in fact change sign between p^0 and p^*. Thus, p' must be outside the interval between p^0 and p^*, which means that

$$(16)\qquad (p'-p^0)(p'-p^*)>0.$$

On the other hand, we know that $(p'-p^0)(p^*-p^0)>0$ from assumption (ii). This result and inequality (16) immediately yield (10).

Proposition 3, we may remark, does not revert trivially to local analysis. It is important to realize that restricting attention to small shifts in θ does not automatically imply small adjustments in p and u. More specifically, if the initial equilibrium is unstable, these adjustments will be large even when the parametric shift is infinitesimal. Thus, our analysis is still global in essence.

III. An Application: Transfers and Welfare

Samuelson (1947) proved that a purchasing-power transfer from donor to recipient never has a paradoxical effect on welfare in the standard two-agent, two-good model if the initial equilibrium is Walras stable. Our Proposition 2 now enables us to establish that his theorem is valid regardless of the magnitude of the transfer and the stability of initial equilibrium.[8]

Let an increase in θ represent a purchasing-power transfer from the donor to the recipient, and have u denote the utility level of the latter agent. Without loss of generality we may assume that, in trading with the donor, the recipient is a net seller of good X.

[8] See also our (1984) Theorem 3 and our related discussion of Yves Balasko (1978). Recall that the analysis of the transfer problem was the fertile ground for Samuelson (1971) and us (in our original version of the present paper) to raise the global issues discussed here.

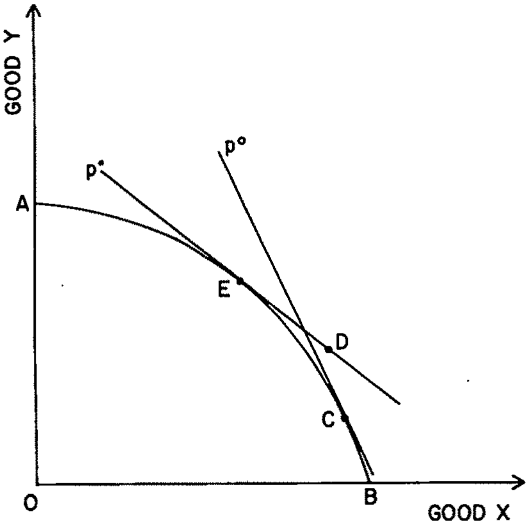


FIGURE 2

Noting that the primary impact of the transfer upon u is clearly positive, we suppose that the secondary impact is negative, since otherwise the possibility of a paradox would be trivally precluded.

In Figure 2, curve AB is the production-possibility frontier for the two-agent economy as a whole; point C is the initial equilibrium for aggregate production and consumption at price p^0; while points E and D represent aggregate production and consumption, respectively, if the recipient's terms of trade were to deteriorate to p^*. (At this stage in the argument, D may be located anywhere on the p^* line.) Since convexity of the production-possibility set ensures that C lies below the p^* line, a well-known theorem due to John Hicks (1940)[9] implies that no distribution of bundle C can make both agents as well off as with the actual distribution of bundle D. Thus, recalling that $v(p^*,\theta')$ at D equals $v(p^0,\theta^0)$ at C (by definition of p^* as the welfare-preserving price in the face of a transfer), we immediately see that the donor's welfare must be greater at D than at C. In light of this information, reapplication of the Hicks theorem implies that D cannot lie below the p^0 line. Therefore, noting that E does lie below this line (by convexity once again), we conclude that D must be located southeast of E. In other words, the necessary condition (13) for a welfare paradox is not satisfied, since $(p' - p^0) < 0$ and $x(p^*,\theta') > 0$.

Hence, Proposition 2 implies that the transfer's overall effect on welfare is normal.[10] This result, moreover, holds for a transfer of any magnitude, regardless of whether the initial equilibrium is stable.

IV. Concluding Remarks

The present paper has generalized the Global Correspondence Principle by extending, in two major ways, Samuelson's 1971 analysis of the exchange-rate response to an international purchasing-power transfer. First, we analyzed the price effect of a parametric shift by allowing the shift to occur in any parameter, not necessarily a transfer payment. Second, we explored the resulting adjustments in any nonprice variable such as welfare. As our analysis showed, the direction of these adjustments is independent of both their magnitude (small or large) and the local nature of equilibrium (stable or unstable). Thus, we have generalized the conventional algebra of comparative statics, which typically assumes small shifts from a stable equilibrium.

What about "higher dimensionality"? Our generalization is by no means limited to two agents. In fact, Proposition 1 is really about the aggregate economy, without reference to the number of agents included; while Propositions 2 and 3 focus on one agent as distinct from the rest of the aggregate economy, whose underlying composition is not essential to the analysis.[11]

[9] As Hicks states: "Thus if we start from any actual distribution of wealth in the I situation, what the condition $\Sigma p_2 q_2 > \Sigma p_2 q_1$ tells us is that it is impossible to reach, by redistribution, a position in which everyone is as well off as he is in the II situation" (p. 111).

[10] This result need not hold if there were tax distortions, as shown by Brecher and Bhagwati (1982) and our paper (1985).

[11] The specific analysis of Figure 2, however, is restricted to the two-agent case, for reasons suggested by, for example, our paper (1983) and references cited therein.

Generalization to more than two goods, however, is another matter,[12] as noted also by Samuelson (1971). Under what restrictions such a generalization may be possible in the many-good case is an interesting question for further research.

[12] In the two-good case, the tâtonnement process is always stable if the equilibria are isolated. This special property, however, does not hold in general when we introduce additional goods. See Kenneth Arrow and F.H. Hahn (1971, pp. 282–85).

REFERENCES

Arrow, Kenneth J. and Hahn, F. H., *General Competitive Analysis*, San Francisco: Holden-Day, 1971.

Balasko, Yves, "The Transfer Problem and the Theory of Regular Economies," *International Economic Review*, October 1978, *19*, 687–94.

Bhagwati, Jagdish N., "Immiserizing Growth: A Diagrammatic Analysis," *Review of Economic Studies*, June 1958, *25*, 201–05.

______, **Brecher, Richard A. and Hatta, Tatsuo,** "The Generalized Theory of Transfers and Welfare: Bilateral Transfers in a Multilateral World," *American Economic Review*, September 1983, *73*, 606–18.

______, ______, **and** ______, "The Paradoxes of Immiserizing Growth and Donor-Enriching (Recipient-Immiserizing) Transfers: A Tale of Two Literatures," *Weltwirtschaftliches Archiv*, No. 2, 1984, *120*, 228–42.

______, ______, **and** ______, "The Generalized Theory of Transfers and Welfare: Exogenous (Policy-Imposed) and Endogenous (Transfer-Induced) Distortions," *Quarterly Journal of Economics*, August 1985, *100*, 697–714.

Brecher, Richard A. and Bhagwati, Jagdish N., "Immiserizing Transfers from Abroad," *Journal of International Economics*, November 1982, *13*, 353–64.

Hicks, John R., "The Valuation of Social Income," *Economica*, May 1940, *7*, 105–24.

Samuelson, Paul A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947; enlarged ed., 1983.

______, "On the Trail of Conventional Beliefs about the Transfer Problem," in Jagdish N. Bhagwati et al., eds., *Trade, Balance of Payments and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*, Amsterdam: North-Holland, 1971, 327–51.

Using Survey Data to Test Standard Propositions Regarding Exchange Rate Expectations

By JEFFREY A. FRANKEL AND KENNETH A. FROOT*

Survey data provide a measure of exchange rate expectations superior to the forward rate in that no risk premium interferes. We estimate extrapolative, adaptive, and regressive models of expectations. Static or "random walk" expectations and bandwagon expectations are rejected: current appreciation generates the expectation of future depreciation because variables other than the contemporaneous spot rate receive weight. In comparing expectations to the process governing the spot rate, we find statistically significant bias.

No variable is as ubiquitous in international financial theory and yet as elusive empirically as investors' expectations regarding exchange rates. In the past, expectations have been modeled in an *ad hoc* way, often by using the forward exchange rate. There is, however, a serious problem with using the forward discount as the measure of the expected change in the exchange rate, in that the two may not be equal. The gap that may separate the forward discount and expected depreciation is generally interpreted as a risk premium. Most of the large empirical literature testing the unbiasedness of the forward exchange rate, for example, has found it necessary either arbitrarily to assume away the existence of the risk premium, if the aim is to test whether investors have rational expectations, or else to assume that expectations are in fact rational, if the aim is to test propositions regarding the behavior of the risk premium.

We offer a new source of data to measure exchange rate expectations that avoids such problems: three independent surveys of the expectations held by exchange market participants. Between 1976 and 1985, American Express Banking Corporation (Amex) polled a sample of 250–300 central bankers, private bankers, corporate treasurers, and economists, regarding their expectations of major exchange rates six months and twelve months into the future, approximately once a year. Since 1981, the *Economist Financial Report*, a newsletter associated with the *Economist*, has conducted at regular six-week intervals a survey of fourteen leading international banks regarding their expectations at three, six, and twelve-month horizons. And since 1983, Money Market Services, Inc. (MMS) has conducted a similar survey on a weekly or biweekly basis, at a variety of short-term horizons. The first two surveys record expectations of five currencies against the dollar (the pound, French franc, mark, Swiss franc, and yen), and the MMS data have been collected for four currencies (the pound, mark, Swiss franc, and yen). In each survey, it is the median response that is reported.

In this paper we are interested principally in two questions: how best to describe the survey expectations in terms of simple models of investors' expectations formation; and whether investors' expectations are unbiased forecasts of the actual spot exchange rate process. Our aim here is not to develop any special new hypotheses of our own. But a

*Department of Economics, University of California, Berkeley, CA 94720, and Sloan School of Management, MIT, Cambridge, MA 02139, respectively. This is an extensively revised version of NBER Working Paper No. 1672. We thank Barbara Bruer, John Calverly, Louise Cordova, Kathryn Dominguez, Laura Knoy, Stephen Marris, and Money Market Services, Inc. for help in obtaining data, the National Science Foundation (under grant no. SES-8218300), the Institute for Business and Economic Research at UC-Berkeley, and the Alfred P. Sloan Foundation's doctoral dissertation program for research support. The data used in this paper are available on request from the authors.

theme which runs throughout our investigation is the stability of expectations. Do the data confirm the suspicions of some critics of floating exchange rates that expectations are characterized by bandwagon effects? Or, in line with many macro models of exchange rate determination, does a current appreciation of the currency by itself generate expectations of future depreciation?

The paper is organized as follows. Section I discusses the exchange rate survey data. In Section II, we present some simple but enlightening summary statistics from the surveys. In Section III, we attempt to describe the survey data by using several popular formulations for exchange rate expectations: extrapolative, adaptive, and regressive models. Section IV then investigates the behavior of the actual spot process and the rationality of the various expectations mechanisms considered in Section III. In Section V, we offer some thoughts on heterogeneity of exchange rate expectations, and Section VI gives our conclusions.

I. The Survey Data

Economists generally distrust survey data. It is a cornerstone of "positive economics" that we learn more by observing what people do in the marketplace than what they say. Nevertheless, alternative measures of expectations all have their own drawbacks. For this reason, closed-economy macro and financial economists have found survey data useful, in studies of expected inflation (where the Livingston survey has been the most popular), expected official announcements of the money stock and other macroeconomic variables (where MMS is the source), and firm inventory behavior and related topics (see Michael Lovell, 1986). To our knowledge, there had been no studies prior to this one using survey data on exchange rate expectations.[1] This might be considered surprising in light of the great interest in the subject, evident in the large literature on the forward market. One could even argue that the case for using survey data on exchange rate expectations is on firmer ground than the case for using survey data on inflation expectations. The respondents to the surveys participate more directly in the spot and forward exchange markets than the respondents to the Livingston survey participate in the goods markets: they are economists in the foreign exchange trading room or the traders themselves in major international banks who have up-to-the-minute information on the values of the currencies covered. At the very least, these exchange rate survey data contain some useful information that warrants study. It seems likely that economists have not used the data in the past only because they have been unaware of their existence.

One limitation to the survey data should be registered from the start, the relatively small number of times the surveys were conducted as of early 1986: 12 dates for the Amex data, 38 for the *Economist* data, 47 for the 1983–84 MMS survey.[2] By pooling the cross section of four or five currencies at each survey date, however, we achieve respectable sample sizes. The obvious contemporaneous correlation of error terms across currencies may be exploited, and we do so with two techniques. Seemingly unrelated regressions are used in cases where the error terms are serially uncorrelated, while method of moments estimators are employed when under the null hypothesis there is serial correlation.[3] In addition, there is considerable

[1] Richard Levich (1979) studies the predictions of the exchange rate forecasting industry. For a recent study of exchange rate expectations using the MMS survey data, see Kathryn Dominguez (1986).

[2] A second limitation of the Amex survey is that it is conducted by mail, and therefore precise dating of expectations was impossible. In response to this problem, we used several alternative methods of dating in all our tests. It turned out that the dating method had a negligible effect on the results. See the Data Appendix for more detail.

[3] In the NBER working paper version of this paper, we also estimated bootstrap standard errors, which are robust in small samples, with respect to estimators that are nonlinear in the residuals and with respect to a variety of nonnormal distributions. This technique has been omitted here both because the resulting standard errors were not very different from those obtained using more conventional methods and because we now have

variety of forecast horizon in the data we employ. We estimate equations for the pooled data at three-, six-, and twelve-month horizons for the *Economist* data, three-months for the MMS data, and six- and twelve-months for the Amex data.

II. Preliminary Results

Before we set out to test the hypotheses of interest, some descriptive statistics and preliminary tests are in order.

A. *The Magnitude of Expected Depreciation*

First, the survey data can be used to shed some light on questions concerning the size of expected depreciation relative to the forward discount. In general, the forward discount can be decomposed into expected depreciation and the risk premium:

$$fd_t = \Delta s^e_{t+1} + rp_t,$$

where fd_t is the log of the forward rate minus the log of the spot rate at time t (expressed in dollars per unit of foreign currency), and Δs^e_{t+1} is the log of the expected future spot rate minus the log of the current spot rate. Many models of exchange rate determination have made the simplifying (but extreme) assumption that expectations are static, for lack of a better alternative, that is, that expected depreciation is zero:

$$(1) \qquad \Delta s^e_{t+1} = 0.$$

For example, William Branson, Hannu Halttunen, and Paul Masson did so, giving as a reason that "we have very little empirical evidence on alternative, more complicated expectations mechanisms" (1977, p. 308). The immortal Mundell-Fleming model of exchange rates under conditions of perfect capital mobility can be interpreted as having assumed static expectations, so that international arbitrage equated domestic and foreign interest rates.

More recently, this point of view has been, in a sense, vindicated by the work of Richard Meese and Kenneth Rogoff (1983). They have shown that the current spot exchange rate is a better predictor of the future rate than are standard monetary models, more elaborate time-series models, or the current forward exchange rate; that is, that the exchange rate seems to follow a random walk. Similar empirical findings have turned up in other contexts. Many papers, such as John Bilson (1981) and Roger Huang (1984), have reported evidence that the rational expectation is closer to zero depreciation than to the forward discount. These authors did not explicitly conclude that the same is necessarily true of investors' expectations; they found support for the random walk model of the spot rate, but were relatively agnostic on investors' expectations.

Nevertheless, this work seems to imply that investors' expected depreciation is not a very interesting variable—that it does not differ very much from zero and is not very responsive to changes in the contemporaneous information set. Bilson (1985) seems to express this point of view, holding that "actual or market forecasts of exchange rates" are unrelated to the forward discount. The position in the Bilson paper is, in effect, that the random walk holds not only as a description of the actual spot rate process but also as a description of investors' expectations formation. It follows that the risk premium constitutes the entire forward discount.

A very different impression of the relative importance of expected depreciation as a component of the forward discount is given by all three of our surveys. Table 1a shows, for each of the surveys, expected depreciation of the dollar against all currencies for which data are available. Most striking is that the survey-expected depreciation is not only consistently positive, but is larger (often several times larger) than the expected depreciation implied by the contemporaneous forward discounts reported in Table 1b. An important feature of Table 1a is the apparent agreement across different surveys and forecast horizons. The corroboration of such large expected depreciation numbers sug-

several times as many observations for the *Economist* data and have added the MMS sample to the analysis.

TABLE 1a—SURVEY EXPECTED DEPRECIATION OF THE DOLLAR AGAINST FIVE CURRENCIES

Data Set	1976–79	1981	1982	1983	1984	1985
MMS 3-Month				8.17	7.26	
Economist 3-Month		9.95	13.44	10.17	10.68	1.56
Economist 6-Month		8.90	10.31	10.42	11.66	3.93
Amex 6-Month	1.20	7.60	10.39	4.19	9.93	1.16
Economist 12-Month		7.17	8.33	7.65	10.02	4.24
Amex 12-Month	−0.20	5.67	6.86	5.18	8.47	3.60

Note: MMS data are the average of four currencies (the pound, mark, Swiss franc, and yen) and do not include the French franc.

TABLE 1b—FORWARD DISCOUNT OF THE DOLLAR AGAINST FIVE CURRENCIES

Time Sample	1976–79	1981	1982	1983	1984	1985
MMS 3-Month				3.05	4.60	
Economist 3-Month		3.94	2.95	1.17	3.20	1.22
Economist 6-Month		3.74	3.01	1.10	3.21	0.84
Amex 6-Month	1.06	4.49	5.21	1.48	4.39	0.02
Economist 12-Month		3.40	3.02	1.25	3.29	0.89
Amex 12-Month	0.93	3.70	4.65	1.28	4.45	0.31

Notes: Forward discounts were recorded at the time each survey was conducted. See the Data Appendix for more detail. MMS data are the average of four currencies (the pound, mark, Swiss franc, and yen) and do not include the French franc.

gests that the results are not due to the particularities of each survey's respondents. Table 2 shows the averages of alternative measures of expected depreciation by survey and by country. The forward discount numbers seem to imply that, on average, the dollar was expected to depreciate against the mark, Swiss franc, and yen, to remain approximately unchanged against the pound, and to appreciate against the franc. The survey expectations, on the other hand, suggest that the results in Table 1a do not mask a great deal of variation across countries. Table 2 shows that the surveys consistently predicted substantial depreciation of the dollar against all five currencies surveyed. In every survey, expected depreciation is considerably smaller, however, for currencies that were selling forward at a smaller discount (or a larger premium).

These simple results provide some indication that market expectations are positively correlated, at least cross sectionally, with the forward discount. Such systematic relationships between expected depreciation and other contemporaneous variables suggest that there is more to investor expectations than is revealed by the random walk model of expectations.[4]

B. *Unconditional Bias*

The simplest possible test of rational expectations is to see if expectations are unconditionally biased, if investors systematically overpredict or underpredict the future spot rate. Tests performed in the 1970's clearly failed to find any unconditional bias.[5] But in the 1980's, the dollar has consistently sold at a discount in the forward exchange market against the most important currencies, as is shown in Tables 1b and 2, and yet it was not until 1985 that the great, long-anticipated dollar depreciation began to materialize. Indeed, George Evans (1986)

[4] Froot and Frankel (1986) decompose the *variance* of the forward discount into expected depreciation and the risk premium. In the present paper we are concerned only with the first moments.

[5] See Bradford Cornell (1977), Alan Stockman (1978), and Frankel (1980).

TABLE 2—VARIOUS MEASURES OF EXPECTED DEPRECIATION OVER THE FOLLOWING MONTHS
(Percent per annum)

			Survey Data		Forward Discount		Actual Change
Forecast Horizon	Survey Source	Dates	N	$E[s(t+1)]-s(t)$	$f(t)-s(t)$	N	$s(t+1)-s(t)$
1 Week							
Total	MMS	10/84–2/86	247	1.03		247	20.20
UK			62	−12.84		62	14.96
WG			62	2.84		62	21.36
SW			61	8.84		61	20.10
JA			62	5.40		62	24.39
2 Weeks							
Total	MMS	1/83–10/84	187	4.22		187	−12.35
UK			47	−2.66		47	−16.15
WG			47	5.09		47	−15.19
SW			46	6.10		46	−13.86
JA			47	8.40		47	−4.23
1 Month							
Total	MMS	10/84–2/86	176	−2.63	1.23	176	20.82
UK			44	−11.91	−3.85	44	10.13
WG			44	−2.26	3.23	44	23.82
SW			44	0.67	3.74	44	21.76
JA			44	2.99	1.68	44	27.55
3 Months							
Total	MMS	1/83–10/84	187	7.76	3.75	187	−10.77
UK			47	4.46	0.37	47	−13.92
WG			47	8.33	4.68	47	−13.68
SW			46	9.62	6.13	47	−12.61
JA			47	8.68	3.85	47	−2.90
Total	*Economist*	6/81–12/85	190	9.13	2.20	195	−0.84
UK			38	3.66	−0.06	38	−6.43
FR			38	5.17	−3.94	38	−4.43
WG			38	11.84	4.36	38	0.81
SW			38	12.30	5.99	38	1.47
JA			38	12.66	4.67	38	4.37
6 Months							
Total	*Economist*	6/81–12/85	190	9.30	2.22	180	−2.18
UK			38	4.19	0.14	36	−6.79
FR			38	4.69	−4.03	36	−6.29
WG			38	12.39	4.35	36	−0.96
SW			38	12.27	5.89	36	−0.36
JA			38	12.94	4.74	36	3.52
Total	Amex	1/76–8/85	51	3.87	2.07	51	5.98
Early Period		1/76–12/78	26	1.20	1.06	26	8.98
Later Period		6/81–8/85	25	6.66	3.12	25	2.86
12 Months							
Total	*Economist*	6/81–12/85	195	7.77	2.31	155	−6.42
UK			38	3.38	0.36	31	−9.47
FR			38	3.72	−3.63	31	−11.20
WG			38	10.67	4.24	31	−5.60
SW			38	10.41	5.91	31	−5.75
JA			38	10.67	4.66	31	−0.08
Total	Amex	1/76–8/85	51	2.81	1.88	46	2.02
Early Period		1/76–12/78	26	−0.20	0.93	26	8.85
Later Period		6/81–8/85	25	5.95	2.88	20	−6.86

TABLE 3—UNCONDITIONAL BIAS IN PREDICTIONS OF FUTURE EXCHANGE RATES
(Percent per annum)

				Survey Error $s^e(t+1)-s(t+1)$			Forward Discount Error $f(t)-s(t+1)$		
Forecast Horizon	Survey Source	Dates	*N*	Mean	*SD* of Mean	*t*-Statistic	Mean	*SD* of Mean	*t*-Statistic
1 Week									
Total	MMS	10/84–2/86	247	−19.17	8.17	−2.35			
UK			62	−27.79	19.87	−1.40			
WG			62	−18.52	15.25	−1.21			
SW			61	−11.27	17.82	−0.63			
JA			62	−18.99	10.97	−1.73			
2 Weeks									
Total	MMS	1/83–10/84	187	16.57	3.37	4.92			
UK			47	13.49	6.70	2.01			
WG			47	20.28	7.43	2.73			
SW			46	19.95	6.42	3.11			
JA			47	12.63	6.25	2.02			
1 Month									
Total	MMS	10/84–2/86	176	−23.44	6.78	−3.46	−19.59	6.31	−3.10
UK			44	−22.04	15.19	−1.45	−13.98	13.26	−1.05
WG			44	−26.08	12.62	−2.07	−20.59	11.77	−1.75
SW			44	−21.09	13.96	−1.51	−18.02	13.12	−1.37
JA			44	−24.57	12.27	−2.00	−25.88	12.10	−2.14
3 Months									
Total	MMS	1/83–10/84	187	18.53	2.88	6.44	14.51	2.86	5.08
UK			47	18.38	5.91	3.11	14.29	5.90	2.42
WG			47	22.01	5.89	3.73	18.36	5.99	3.07
SW			46	22.23	5.20	4.28	18.74	4.85	3.86
JA			47	11.58	5.14	2.25	6.75	4.97	1.36
Total	*Economist*	6/81–12/85	190	9.97	2.92	3.42	3.04	2.73	1.12
UK			38	10.09	6.66	1.51	6.37	5.88	1.08
FR			38	9.61	6.47	1.48	0.49	5.98	0.08
WG			38	11.02	6.45	1.71	3.55	5.90	0.60
SW			38	10.83	7.03	1.54	4.52	6.73	0.67
JA			38	8.29	5.95	1.39	0.30	5.84	0.05
6 Months									
Total	*Economist*	6/81–12/85	180	11.70	3.20	3.66	4.48	3.03	1.48
UK			36	11.32	6.71	1.69	7.10	6.24	1.14
FR			36	11.08	7.13	1.55	2.15	6.71	0.32
WG			36	13.56	7.16	1.89	5.36	6.63	0.81
SW			36	12.77	7.80	1.64	6.37	7.37	0.86
JA			36	9.76	6.84	1.43	1.41	6.65	0.21
Total	Amex	1/76–8/85	51	−2.11	2.82	−0.75	−3.92	2.61	−1.50
Early Period		6/76–12/78	26	−7.78	2.94	−2.65	−7.93	2.80	−2.83
Later Period		6/81–8/85	25	3.79	4.59	0.83	0.26	4.30	0.06
12 Months									
Total	*Economist*	6/81–12/85	155	14.83	2.23	6.64	9.00	2.39	3.77
UK			31	13.73	4.96	2.77	10.39	5.46	1.90
FR			31	15.10	4.75	3.18	7.20	5.09	1.41
WG			31	17.02	4.72	3.60	10.02	4.82	2.08
SW			31	16.73	5.06	3.31	12.13	5.41	2.24
JA			31	11.59	5.02	2.31	5.15	5.27	0.98
Total	Amex	1/76–8/84	46	0.71	2.52	0.28	0.04	2.30	0.02
Early Period		6/76–12/78	26	−9.05	3.20	−2.83	−7.92	3.36	−2.36
Later Period		6/81–8/84	20	13.40	1.07	12.52	10.38	1.10	9.42

Note: Degrees of freedom used to estimate standard deviation (*SD*) of the mean are the number of nonoverlapping observations for each data set.

uses a nonparametric sign test on the forward rate prediction errors over the 1981–84 period and finds significant unconditional bias against the pound. Could there be unconditional bias in the survey data for this period as well?

Table 3 reports formal tests of unconditional bias. The MMS three-month data, available for the period January 1983 to October 1984, show statistically significant bias for all four currencies, even more than the three-month forward discount data during the same period. The *Economist* data are available through 1985, the first year of dollar decline. The bias is not quite statistically significant at the three- and six-month horizons, but it is significant at the one-year horizon.[6] The general rule seems to be that when the forward discount is biased, the survey data are also biased, with the implication that the finding cannot be attributed to a risk premium. The presence of biasedness in the 1980's clearly arises from the episode of dollar appreciation that ended in February 1985. Respondents consistently overpredicted the future value of foreign currencies against the dollar in this period.

One explanation that could be suggested for such findings of biasedness is that the surveys measure investors' expectations with error. But it should be noted that if one is willing to assume that the measurement error is random, then the conclusions are unaffected. Under the null hypothesis, positive and negative measurement errors should average out, just like positive and negative prediction errors by investors.

Short of concluding that investors' expectations are not equal to the rationally expected value, one major possible explanation for findings of biasedness remains. It is that the standard errors in our tests are invalidated by the "peso problem" of nonnormality in the distribution of the test statistic. The peso problem arises when there is a small probability of a large change in the exchange rate each period—such as results from a devaluation, a bursting of a speculative bubble, or a big change in fundamentals —and when the sample size is not large enough to invoke the central limit theorem with confidence.[7,8]

The sensitivity of the direction and magnitude of the bias in prediction error is evident in the Amex survey, the only one available in 1976–79. These data show unconditional bias in the opposite direction in the earlier period, as do the forward rate data: respondents consistently underpredicted the value of foreign currencies against the dollar. When the entire Amex data set from 1976 to 1985 is used, prediction errors show no unconditional bias for either the survey data or the forward rate.

[6] For all data sets but the Amex 6-month, prediction errors are overlapping because the surveys are conducted more frequently than the forecast interval. The standard errors reported for each currency in Table 3 reflect the number of nonoverlapping intervals in each data set, and are thus upper bounds. Higher significance levels could be obtained by combining the results for different currencies. But the apparent low standard errors when all observations are simply pooled are misleading, as there is a definite correlation of errors across currencies at any point in time. The proper technique (*SUR*) for this problem is applied in the following section.

[7] Calculations in Frankel (1985) undermine the hypothesis that the forward discount rationally reflected the 1981–85 path of dollar appreciation, even allowing for the possibility of a sudden large collapse in the dollar.

[8] It should be noted that a fourth explanation sometimes given for findings of biasedness in the forward rate, after the existence of a risk premium, a failure of rational expectations and the peso problem, is the convexity term due to Jensen's Inequality (see Charles Engel, 1984). Note, however, that if exchange rates are lognormally distributed this convexity term is bounded above by the unconditional variance of the spot rate and is therefore small. For a lognormally distributed random variable, $X=e^x$, $E[X]=\int e^x f(x)\,dx=\exp[\mu+(1/2)\sigma^2]$ and $E[1/X]=\int e^{-x} f(x)\,dx=\exp[\mu-(1/2)\sigma^2]$, where

$$f(x)=(1/2\pi)\exp\left[-(x-\mu)^2/2\sigma^2\right].$$

Thus, $\log(E[X])-\log(E[1/X])=\sigma^2$, which is weakly greater than the conditional variance, provided that expectations are formed rationally. During the 1980's, $\sigma^2=0.02$ for the spot rate, so that Jensen's Inequality is too small to explain the magnitude of the forward rate prediction errors, let alone the very large shift of about 18 percent between the late 1970's and early 1980's in Table 3.

III. Tests of Expectations Formation

The question of what mechanisms investors use to form expectations is of interest independent of the question of whether these mechanisms are rational, that is, whether they coincide with the mathematical expectation of the actual spot process. In this section we investigate alternative specifications of expectations, and in Section IV we test for their rationality.

A number of simple formulations have traditionally been used. A general framework for expressing them comes from writing the investors' expected future (log) spot rate as a weighted average of the current (log) spot rate with weight $1-\beta$ and some other element, x_t, with weight β:

$$s^e_{t+1}=\beta x_t+(1-\beta)s_t. \tag{2}$$

In examining different versions of equation (2), our null hypothesis will be that expectations are in fact static, that is, that $\beta=0$ (investors believe in the random walk). We choose interesting candidates for the "other element," x_t, as alternative hypotheses. The models we will consider are extrapolative expectations, adaptive expectations, and regressive expectations. They feature as the other element x_t: the lagged spot rate, s_{t-1}, the lagged expectation, s^e_t, and some notion of a long-run equilibrium level of the spot rate, $\bar{s}_t$, respectively.

One characterization of expectations formation often claimed by market participants themselves is that the most recent trend is extrapolated: if the currency has been depreciating, then investors expect that it will continue to depreciate.[9] Such "bandwagon" expectations are represented:

$$\Delta s^e_{t+1}=-g\Delta s_t, \tag{3}$$

where Δs_t is the most recent observed change in the log of the exchange rate and g is hypothesized to be less than zero. (Again, static expectations would be the special case where $g=0$.) It has long been a concern of critics of floating exchange rates that bandwagon expectations would render the system unstable. For example, Ragnar Nurkse states

> [Speculative] anticipations are apt to bring about their own realization. Anticipatory purchases of foreign exchange tend to produce or at any rate to hasten the anticipated fall in the exchange value of the national currency, and the actual fall may set up or strengthen expectations of a further fall.... Exchange rates under such circumstances are bound to become highly unstable, and the influence of psychological factors may at times be overwhelming. [1944, p. 118]

Nurkse's view was challenged by Milton Friedman (1953), who argued that speculation would be stabilizing. "Speculation" can be defined as buying and selling of currency in response to expectations of exchange rate changes, as compared to the counterfactual case of static expectations. A property of bandwagon expectations is that the expected future spot rate as a function of the observed current spot rate has an elasticity that exceeds unity, as contrasted to static expectations, in which the elasticity is equal to unity. Because investors sell a currency that they expect to depreciate, it follows that under bandwagon expectations, speculation is destabilizing.

The remaining three models we discuss go in the opposite direction. They can all be subsumed under the label *inelastic*, or stabilizing, *expectations*: a change in the current spot rate induces a revision in the expected future level of the spot rate that, though it may be positive, is less than proportionate. An observed appreciation of the currency generates an anticipation of a future depreciation of the currency back, at least partway, toward its previously expected level. If speculators act on the basis of the expected future depreciation, they will put downward pressure on the price of the currency today; in other words, speculation will be stabilizing. One case of inelastic expectations is equation (3) with g greater than zero. An

[9] See, for example, the discussion in Michael Dooley and Jeffrey Shafer (1983, pp. 47–48).

TABLE 4—EXTRAPOLATIVE EXPECTATIONS
(Independent variable: $s(t-1)-s(t)$)

Seemingly Unrelated Regressions[a] of Survey Expected Depreciation:
$s^e(t+1)-s(t)=a+g(s(t-1)-s(t))$

Data Set	Dates	g[c]	$D\text{-}W$[b]	DF	t: $g=0$	R^2
Economist 3-Month	6/81–12/85	0.0416 (0.0210)	1.81	184	1.98[d]	0.30
with AR(1) Correction		0.0463 (0.0195)		179	2.37[e]	0.38
MMS 3-Month	1/83–10/84	−0.0391 (0.0168)	1.49	179	−2.32[e]	0.37
with AR(1) Correction		−0.0298 (0.0203)		194	−1.46	0.19
Economist 6-Month	6/81–12/85	0.0730 (0.0225)	1.36	184	3.25[f]	0.54
with AR(1) Correction		0.0832 (0.0236)		179	3.53[f]	0.58
Amex 6-Month	1/76–8/85	0.2994 (0.0487)	1.89	45	6.15[f]	0.81
Economist 12-Month	6/81–12/85	0.2018 (0.0296)	1.47	184	6.82[f]	0.84
with AR(1) Correction		0.2638 (0.0251)		179	10.51[f]	0.92
Amex 12-Month	1/76–8/85	0.3796 (0.0798)	0.94	45	4.76[f]	0.72

Notes: Asymptotic standard errors are shown in parentheses.
[a] Amex 6 and 12 Month regressions use *OLS* due to the small number of degrees of freedom.
[b] The *D-W* statistic is the average of the equation-by-equation *OLS* Durbin-Watson statistics for each data set.
[c] All equations are estimated allowing each currency its own constant term. To conserve space, estimates of these constant terms are omitted here, but are reported in our papers (1986).
[d] Significant at the 10 percent level.
[e] Significant at the 5 percent level.
[f] Significant at the 1 percent level.

equivalent representation would be

$$s^e_{t+1}=(1-g)s_t+gs_{t-1}, \tag{4}$$

where s_t is the logarithm of the current spot rate and g is hypothesized to be positive. The hypothesis is a simple form of *distributed lag expectations*. Obviously we could have longer lags as well.

In Tables 4–11, we can interpret the regression error as random measurement error in the survey data. Under the joint hypothesis that the mechanism of expectations formation is specified correctly and that the measurement error is random, the parameter estimates are consistent. It should be noted that this joint hypothesis is particularly restrictive because the spot rate appears on the right-hand side; if a change in expected depreciation feeds back to affect both the contemporaneous spot rate and any element of the regression error, then the parameter estimates will be biased and inconsistent. Such simultaneous equation bias, however, is not a problem under our null hypothesis that expected depreciation is constant.

Table 4 reports the results of the Seemingly Unrelated Regressions (SUR)[10] of the survey expected depreciation on the recent change in the spot rate, equation (3), which we call under the general title of extrapolative expectations, where $g>0$ represents the case of distributed lag and $g<0$ represents

[10] Due to the small number of observations in the Amex data sets, *OLS* rather than *SUR* was used to conserve degrees of freedom in this case.

the case of bandwagon expectations.[11] Most of the slope parameters in the column labelled g in Table 4 are positive and significant at the 1 percent level. The evidence suggests that expectations are less than unit elastic with respect to the lagged spot rate, that is, expectations are stabilizing. For example, the point estimate of 0.04 in the three-month *Economist* data set implies an appreciation of 10 percent today generates an expectation of a 0.4 percent depreciation over the next three months, a rate of 1.6 percent per year.

The Durbin-Watson (*D-W*) tests for serial correlation reported in Table 4 (except those for the Amex data sets) are the averages of the equation-by-equation *OLS* regressions used in the first step of the *SUR* procedure. For this reason, and since the Amex data are irregularly spaced and thus are not true time-series, values of the *D-W* test must be interpreted with caution. Nevertheless, the null hypothesis of no "serial" correlation is still appropriate, and the low reported values of the statistic suggest that the standard errors are suspect. To correct for serial correlation in the residuals, we used a generalized three-stage least squares estimator that allows for contemporaneous as well as first-order serial correlation of each country's residual.[12] These results for the *Economist* and MMS data sets are reported beneath the uncorrected *SUR* estimates in Table 4.[13] While we find some evidence of serial correlation in the data, the corrected coefficients are similar in size, and the standard errors are even more unfavorable to the bandwagon hypothesis than in the uncorrected seemingly unrelated regressions. The lone case of a negative point estimate for g, in the three-month MMS sample, loses its statistical significance under the correction for serial correlation.

Despite the rejection of bandwagon expectations in favor of the stabilizing distributed lag, it may still be true that psychological factors are important in foreign exchange markets. The absence of bandwagon effects in the data does not rule out the possibility of speculative bubbles. For example, rational bubbles which are constantly forming and popping would not yield systematic bandwagon effects in the spot rate.

Adaptive expectations are an old standby in the economist's arsenal of expectations models. The expected future spot rate is formed adaptively, as a weighted average of the current observed spot rate and the lagged expected rate:

$$s^e_{t+1} = (1-\gamma_1)s_t + \gamma_1 s^e_t, \tag{5}$$

where γ_1 is hypothesized between 0 and 1 for expectations to be inelastic.[14]

We report the results of regressing expected depreciation on the lagged survey prediction error in Table 5:

$$\Delta s^e_{t+1} = \gamma_1(s^e_t - s_t). \tag{5'}$$

Three of the six coefficients in the column labelled γ_1 are statistically significant. All three are positive, implying that expectations place positive weight on the previous prediction. The results in Table 5 provide evidence in favor of the hypothesis that expectations are stabilizing.[15] The *D-W* statistics are again very low, particularly in the twelve-month data. When we use the three-stage least squares correction for serial correlation, the coefficient is significant in three out of four data sets.

The *regressive expectations* model was made popular by Rudiger Dornbusch

[11] We take the definition of extrapolative expectations from Jacob Mincer (1969).

[12] See R. W. Parks (1967).

[13] Because of irregular spacing, we could not correct the estimates for serial correlation in the Amex data sets.

[14] Adaptive expectations have been considered by Pentti Kouri (1976), as a third alternative after static and rational expectations, as well as by Rudiger Dornbusch (1976a) and many other authors.

[15] An implication of any measurement error in the survey data is that the lagged prediction errors, which appear as regressors in Table 5, are also measured with error. Thus we would expect the point estimates of γ_1 to be biased toward zero. However, in view of the fact that the variance of actual spot rate changes is about 10 times larger than the variance of the survey-expected depreciation (Froot-Frankel, Table 3), we suspect that this bias is small.

TABLE 5—ADAPTIVE EXPECTATIONS
(Independent variable: $s^e(t)-s(t)$)

Seemingly Unrelated Regressions[a] of Survey Expected Depreciation:
$E[s(t+1)]-s(t)=a+\gamma_1(s^e(t)-s(t))$

Data Set	Dates	γ_1[c]	D-W[b]	DF	t: $\gamma_1=0$	R^2
Economist 3-Month	6/81–12/85	0.0798 (0.0203)	2.01	169	3.93[f]	0.63
with AR(1) Correction		0.0716 (0.0180)		164	3.97[f]	0.64
MMS 3-Month	1/83–10/84	−0.0272 (0.0215)	1.29	159	−1.26	0.15
with AR(1) Correction		−0.0234 (0.0234)		154	−1.00	0.10
Economist 6-Month	6/81–12/85	0.0516 (0.0161)	1.12	159	3.20[f]	0.53
with AR(1) Correction		0.0783 (0.0223)		154	3.52[f]	0.58
Amex 6-Month	1/76–8/85	−0.0702 (0.1200)	2.10	15	−0.58	0.04
Economist 12-Month	6/81–12/85	−0.0093 (0.0244)	1.10	139	−0.38	0.02
with AR(1) Correction		0.1890 (0.0301)		134	6.28[f]	0.81
Amex 12-Month	1/76–8/85	0.0946 (0.0212)	0.55	31	4.47[f]	0.69

Notes and footnotes: See Table 4.

(1976b). It is a more elegant specification, consistent with dynamic models in which variables such as goods prices converge toward their long-run equilibrium values over time in accordance with differential equations, or, in discrete time, in accordance with difference equations:

$$s^e_{t+1}=(1-\vartheta)s_t+\vartheta\bar{s}_t. \tag{6}$$

Here $\bar{s}_t$ is the long-run equilibrium exchange rate, and ϑ (a number between 0 and 1 in this discrete-time version) is the speed at which s_t is expected to regress toward $\bar{s}_t$, as can perhaps be seen more clearly in the equivalent representation,

$$\Delta s^e_{t+1}=-\vartheta(s_t-\bar{s}_t). \tag{7}$$

The long-run equilibrium $\bar{s}_t$ can itself change. It is often assumed to obey purchasing power parity, increasing proportionately in response to a change in the domestic money supply and price level.

In the econometric tests below, we try out two alternative formulations for $\bar{s}_t$. The simplest possible description of the long-run equilibrium is that it is constant over our sample. Thus we regress expected depreciation on the spot rate and constant terms for each country. The results are presented in Table 6. A second specification for the long-run value of the exchange rate is that given by purchasing power parity (*PPP*). In this case, $\bar{s}_t$ moves with relative inflation differentials instead of remaining constant:

$$\bar{s}_t=s_0+\log\left(\frac{P_t/P_0}{P_t^*/P_0^*}\right), \tag{8}$$

where s_0 is the log of the average nominal value of the foreign currency in terms of dollars, 1973–79, P_t and P_t^* are the current monthly levels of the U.S. and foreign CPIs, respectively, and P_0 and P_0^* are the average levels of the U.S. and foreign CPIs, 1973–79.

The general conclusions that come out of Tables 6 and 7 are identical. Four of the six data sets give significant weight to the long-run equilibrium, in each case positive. Investors expect the spot rate to regress toward its

TABLE 6—REGRESSIVE EXPECTATIONS I
(Independent variable: $s(t)$; Long-run equilibrium constant)

Seemingly Unrelated Regressions[a] of Survey Expected Depreciation:
$s^e(t+1)-s(t)=a-\theta s(t)$

Data Set	Dates	θ^c	$D\text{-}W^b$	DF	t: $\theta=0$	R^2
Economist 3-Month	6/81–12/85	0.0359 (0.0101)	1.56	184	3.55[f]	0.58
with AR(1) Correction		0.0226 (0.0109)		179	2.07[e]	0.32
MMS 3-Month	1/83–10/84	0.0100 (0.0159)	1.46	179	0.63	0.04
with AR(1) Correction		0.0061 (0.0195)		174	0.31	0.01
Economist 6-Month	6/81–12/85	0.0764 (0.0127)	1.14	184	6.00[f]	0.80
with AR(1) Correction		0.0807 (0.0170)		179	4.73[f]	0.71
Amex 6-Month	1/76–8/85	−0.0000 (0.0235)	1.19	45	−0.00	0.00
Economist 12-Month	6/81–12/85	0.1724 (0.0161)	1.03	184	10.70[f]	0.93
with AR(1) Correction		0.1905 (0.0182)		179	10.48[f]	0.92
Amex 12-Month	1/76–8/85	0.0791 (0.0346)	0.48	45	2.29[e]	0.37

Notes and footnotes: See Table 4.

TABLE 7—REGRESSIVE EXPECTATIONS II
(Independent variable: $\bar{s}(t)-s(t)$; Long-run equilibrium *PPP*)

Seemingly Unrelated Regressions[a] of Survey Expected Depreciation:
$s^e(t+1)-s(t)=a+\theta(\bar{s}(t)-s(t))$

Data Set	Dates	θ^c	$D\text{-}W^b$	DF	t: $\theta=0$	R^2
Economist 3-Month	6/81–12/85	0.0223 (0.0126)	1.66	184	1.78[d]	0.26
with AR(1) Correction		0.0119 (0.0133)		179	0.89	0.08
MMS 3-Month	1/83–10/84	−0.0207 (0.0146)	1.55	179	−1.41	0.18
with AR(1) Correction		0.0083 (0.0194)		174	0.43	0.02
Economist 6-Month	6/81–12/85	0.0600 (0.0159)	1.32	184	3.77[f]	0.61
with AR(1) Correction		0.0782 (0.0221)		179	3.54[f]	0.58
Amex 6-Month	1/76–8/85	0.0315 (0.0202)	1.22	45	1.56	0.21
Economist 12-Month	6/81–12/85	0.1750 (0.0216)	1.25	184	8.10[f]	0.88
with AR(1) Correction		0.2449 (0.0274)		179	8.93[f]	0.90
Amex 12-Month	1/76–8/85	0.1236 (0.0276)	0.60	45	4.48[f]	0.69

Notes and footnotes: See Table 4.

long-run equilibrium. Note that this is a stronger property than the fact, which we discovered in Tables 1a and 2, that investors have been forecasting large depreciation on average throughout the 1980's. Regressivity requires not only that investors expect a currency that is above its long-run level to depreciate, but also that they expect it to depreciate by more the farther it is above its equilibrium value. In Table 7, the *Economist* regressions at three-, six-, and twelve-month horizons show that deviations from *PPP* are expected to decay at annual rates of $(1-0.9881^4)\approx 5$ percent, $(1-0.9218^2)\approx 15$ and 24 percent, respectively. This last figure implies that the expected half-life of *PPP* deviations is 2.5 years.

Clearly, if a high R^2 were our goal, more complicated models could have been reported. We estimated a more general specification for expectations, expanding the information set to include simultaneously the current and lagged spot rates, the long-run equilibrium rate and the lagged expected spot rate. We then tested the entire set of nested hypotheses, beginning with this general specification all the way to static expectations. In particular, we considered as alternatives to the simple models discussed above hybrid specifications such as "adaptive-bandwagon":

$$\Delta s^e_{t+1} = \gamma(s^e_t - s_t) - g\Delta s_{t+1}.$$

The R^2s of these more complex permutations were higher than those reported in Tables 4 through 7. However, the best fits were for models which are unfamiliar compared with the popular formulations above. Furthermore, the strongest statistical rejections were those reported here, of static expectations against the simpler extrapolative, adaptive and regressive models; when estimating the hybrid models, by contrast, we were able statistically to accept the constraints implied by the simple models. For these reasons we do not report the results.

The central point of our analysis is to investigate the robustness of a rejection of static expectations, not to settle on any single model of expectations. The goodness of fit statistics in Tables 4 through 7, however, give us an opportunity to compare the fits of these simple alternative specifications. From this set of alternatives, the best model appears to be the distributed lag.

IV. Are Expectations Formed Rationally?

Now that we have an idea of the parameters describing the formation of investor expectations, we will see how well they correspond to the parameters describing the true process governing the spot rate. We could estimate first the mathematical expectation of the actual spot process conditional on each of the information sets considered in Section III, and only then test for equality with the process governing investors' expectations. Here we report directly regressions of the difference between investor expectations and the realized spot rate ($\Delta s^e_{t+1} - \Delta s_{t+1}$ or, equivalently, $s^e_{t+1} - s_{t+1}$) against the same variables as in the preceding section. Under the null hypothesis the coefficient should be zero, and the error term should be uncorrelated with the right-hand side variables, that is, the spot rate prediction error should be purely random, as should be the case for any right-hand side variables observed at time t. Furthermore, under the null hypothesis, the error term should be serially uncorrelated, which makes the econometrics easier. The logic is the same as in the existing literature of rational expectations tests, where expectations are measured by the forward rate rather than survey data, except that we are free of the problems presented by the risk premium.[16] Because a statistical rejection of the null hypothesis could in theory be due to the failure of the error term to have the proper normal distribution (the peso problem mentioned in Section II, Part B), or could be due to a learning period following a "regime change," rather than to a failure of investors to act rationally, we will use the terms "systematic ex-

[16] In the NBER working paper version, we reported for purposes of comparison in all our tests results both using expectations measured by the forward discount and using expectations measured by the survey data.

pectational errors" or "bias in the sample" to describe the alternative hypothesis, in preference over a "failure of rational expectations."

In testing whether expectations are biased in the sample, there are added advantages in having first tested models of what variables matter for expectations. For those cases in which we fail to reject the null hypothesis, it helps to have an idea whether the right-hand side variable is relevant to determining Δs^e_{t+1} and Δs_{t+1}; if not, the test for the presence of bias is not very powerful. For those cases when we do reject the null hypothesis, we will have a ready-made description of the nature of investors' bias. An explicit alternative hypothesis is lacking in most standard tests.

A. *Econometric Issues*

The tests of rational expectations below were performed by *OLS*, with standard errors calculated using a method of moments procedure. The usual *OLS* standard errors are inappropriate because of the contemporaneous correlation across countries, and a sampling interval many times smaller than the forecast horizon. In the previous section, where expected depreciation is the regressand, a long forecast horizon and short sampling interval do not themselves imply that the error term is serially correlated, since expectations are formed using only contemporaneous and past information. When the prediction error is on the left-hand side, however, we have the usual problem induced by overlapping observations: under the null hypothesis the error term, consisting of new information that becomes available during the forecast interval, is a moving average process of an order equal to the number of sampling intervals contained in the forecast horizon minus one.[17] The *OLS* point estimates remain consistent in spite of the serially correlated residuals. The method of moments estimate of the sample covariance matrix of the *OLS* estimate, $\hat{\beta}$ is

$$(9)\quad \hat{\Theta} = (\mathbf{X}'_{\mathrm{NT}}\mathbf{X}_{\mathrm{NT}})^{-1}\mathbf{X}'_{\mathrm{NT}}\hat{\Omega}\mathbf{X}_{\mathrm{NT}} \times (\mathbf{X}'_{\mathrm{NT}}\mathbf{X}_{\mathrm{NT}})^{-1},$$

where $\mathbf{X}_{\mathrm{NT}}$ is the matrix of regressors of size N (countries) times T (time). The (i, j)th element of the unrestricted covariance matrix, $\hat{\Omega}$ is

$$(10)\quad \hat{\omega}(i,j) = \frac{1}{NT-k}\sum_{l=0}^{N-1}\sum_{t=k+1}^{T}\hat{u}_{t+lT}\hat{u}_{t-k+lT}$$

$$\text{for } mT-n \le k \le mT+n; \quad m=0,\dots,N-1$$

$$= 0 \text{ otherwise,}$$

where n is the order of the MA process, $\hat{u}_{t+lT}$ is the *OLS* residual, and $k=|i-j|$. Such an unrestricted estimate of Ω uses many degrees of freedom; in the case of the *Economist* twelve-month data, $N=5$ and $n=8$, so that the covariance matrix has $N(N+1)n/2$ or 120 independent parameters. We instead estimated a restricted covariance matrix, $\tilde{\Omega}$, with typical element:

$$(11)\quad \tilde{\omega}(t+lT, t-k+pT)$$

$$= \frac{1}{N-1}\sum_{l=0}^{N}\hat{\omega}(t+lT, t-k+pT) \quad \text{if } l=p \text{ and } -n \le k \le n$$

$$= \frac{2}{N(N-1)}\sum_{p=0}^{N-1}\sum_{l=0}^{N-1}\hat{\omega}(t+lt, t-k+pT) \quad \text{if } l \ne p \text{ and } -n \le k \le n$$

$$= 0 \text{ otherwise.}$$

These restrictions have the effect of averaging the own-currency and cross-currency autocorrelation functions of the *OLS* residuals, respectively, bringing the number of independent parameters down to $2n$.

A problem with our estimate of $\tilde{\Omega}$ is that it need not be positive definite in small samples. Whitney Newey and Kenneth West

[17] For the original application of method of moments estimation to exchange rate data with overlapping observations, see Lars Hansen and Robert Hodrick (1980).

TABLE 8—RATIONALITY OF EXTRAPOLATIVE EXPECTATIONS
(Independent variable: $s(t-1)-s(t)$)

OLS Regressions of Survey Prediction Errors: $s^e(t+1)-s(t+1)=a+g(d(t-1)-s(t))$

Data Set	Dates	g	DF	t: $g=0$	R^2	F test $a=0, g=0$
Economist 3-Month	6/81–12/85	0.2501	184	1.48	0.19	1.06
		(0.1695)				
MMS 3-Month	1/83–10/84	−0.2084	182	−1.38	0.18	6.67[c]
		(0.1506)				
Economist 6-Month	6/81–12/85	0.2449	174	0.84	0.07	0.97
		(0.2904)				
Amex 6-Month	1/76–8/85	1.0987	45	2.91[c]	0.48	3.32[c]
		(0.3776)				
Economist 12-Month	6/81–12/85	−0.6516	149	−2.54[b]	0.42	8.09[c]
		(0.2564)				
Amex 12-Month	1/76–8/85	2.0001	40	5.45[c]	0.77	5.28[c]
		(0.3667)				

Notes: All equations are estimated allowing each currency its own constant term. To conserve space, estimates of the constants are omitted here, but are reported in our paper (1986). Methods of Moments standard errors are shown in parentheses.

[a]Significant at the 10 percent level.

[b]Significant at the 5 percent level.

[c]Significant at the 1 percent level.

(1985) offer a consistent estimate of Ω that discounts the jth order autocovariance by $1-(j/(m+1))$, and is positive definite in finite sample. For any given sample size, however, there is still a question of how large m must be to guarantee positive definiteness. In the subsequent regressions we tried $m=n$ (which Newey and West themselves suggest) and $m=2n$; we report standard errors using the latter value of m because they were consistently larger than those using the former.

B. *The Results*

We now turn to the results of our tests of rationality within the three models examined in Section III.

In Table 4, we found that if investors' expected future spot rate is viewed as a distributed lag of the actual spot rate, then the weight on the current spot rate is less than one and the weight on the lagged spot rate greater than zero. Is this degree of inelasticity of expectations rational? Or is the future spot rate more likely to lie in the direction of the current spot rate, as would be the case if the actual spot rate followed a random walk?

Table 8 shows highly significant rejections for three of the six data sets of the hypothesis that expectations exhibit no systematic bias. As in the case of unconditional bias, the results are immune to measurement error in the survey data, provided the error is orthogonal to the regressors. The *Economist* twelve-month data significantly overestimate the tendency for the spot rate to keep moving in the same direction as it had been, while the Amex data *underestimate* the tendency to keep moving in the same direction. The diversity of results is not primarily attributable to a difference between the two surveys. Table 4 showed similar parameters of expectations formation in the two surveys. Rather the difference is primarily attributable to the behavior of the actual spot process during the two different sample periods for which data are available. If one includes in the sample the years 1976–78, during which the Amex data are available, then more extrapolative expectations would have been correct, because the dollar had a long run of declines followed by a long run of

TABLE 9—RATIONALITY OF ADAPTIVE EXPECTATIONS
(Independent variable: $s^e(t)-s(t)$)

OLS Regressions of Survey Prediction Errors: $s^e(t+1)-s(t+1)=a+\gamma(s^e(t)-s(t))$

Data Set	Dates	γ	*DF*	t: $\gamma=0$	R^2	*F* test $a=0,\gamma=0$
Economist 3-Month	6/81–12/85	0.4296 (0.1395)	169	3.08[c]	0.51	3.39[c]
MMS 3-Month	1/83–10/84	−0.2289 (0.2207)	158	−1.04	0.11	6.35[c]
Economist 6-Month	6/81–12/85	0.0884 (0.2488)	149	0.36	0.01	1.52
Amex 6-Month	1/76–8/85	0.5571 (0.5227)	15	1.07	0.11	1.04
Economist 12-Month	6/81–12/85	−1.0310 (0.2452)	109	−4.20[c]	0.66	10.27[c]
Amex 12-Month	1/76–8/85	0.5972 (0.1007)	25	5.93[c]	0.80	8.05[c]

Notes and footnotes: See Table 8.

appreciation. But if one considers the period 1981–85 alone, *less* extrapolative expectations would have been correct, because first differences of the actual spot rate (though usually negative) were not positively serially correlated.[18] The conclusion is that the actual spot process is significantly different from investors' expectations, but it is also more complicated than a simple distributed lag with constant weights, whether correctly perceived by investors or not.

In Table 5, we found that investors' expectations can be viewed as adaptive. When investors make a prediction error, they revise their previous expectations most, though not all, of the way to the new observed spot rate. Would they do better to revise their expectation even farther, or less far? Assume that the true best predictor of the future spot rate is a weighted average of the current spot rate and the lagged expectation:

$$(12)\quad s_{t+1}=(1-\gamma_2)s_t+\gamma_2 s_t^e+\varepsilon_{t+1}.$$

Then investors' expectations would be rational if and only if γ_1 from equation (5) were equal to γ_2 from equation (12). Taking the difference of the two equations,

$$(13)\quad s_{t+1}^e-s_{t+1}=(\gamma_1-\gamma_2)(s_t^e-s_t)+\varepsilon_{t+1}.$$

In Table 9, we regress the expectational error against the lagged expectational error as in equation (13). Such tests of serial correlation are a common way of testing for efficiency in the forward market.[19] In the context of adaptive expectations, we can see clearly what the alternative hypothesis is. Positive serial correlation is precisely the hypothesis that expectations are insufficiently adaptive; investors could avoid making the same error repeatedly if they revised their expectations all the way to the new spot rate. Negative serial correlation is the hypothesis that expectations are overly adaptive. Table 9 shows that expectations are insufficiently adaptive in four of six data sets. In two cases, the tendency for investors to put too little weight on the current spot rate is highly significant statistically. In one case (the *Economist* twelve-month data), investors put too much weight on the current spot rate relative to the weight they place on the lagged

[18] In the NBER working paper version, we report in each table separate regressions for the actual spot process.

[19] See, for example, Dooley-Shafer and Hansen-Hodrick.

TABLE 10—RATIONALITY OF REGRESSIVE EXPECTATIONS I
(Independent variable: $s(t)$; Long-run equilibrium constant)

OLS Regressions of Survey Prediction Errors: $s^e(t+1)-s(t+1)=a-\theta s(t)$

Data Set	Dates	θ	*DF*	t: $\theta=0$	R^2	*F* test $a=0, \theta=0$
Economist 3-Month	6/81–12/85	−0.1686 (0.0934)	184	−1.80[a]	0.27	1.20
MMS 3-Month	1/83–10/84	−0.0288 (0.1431)	182	−0.20	0.00	6.02[c]
Economist 6-Month	6/81–12/85	−0.3582 (0.1936)	174	−1.85[a]	0.28	1.40
Amex 6-Month	1/76–8/85	−0.0427 (0.1647)	45	−0.26	0.01	2.07[a]
Economist 12-Month	6/81–12/85	−0.4167 (0.1895)	149	−2.20[c]	0.35	6.54[c]
Amex 12-Month	1/76–8/85	0.1904 (0.2919)	40	0.65	0.05	0.36

Notes and footnotes: See Table 8.

TABLE 11—RATIONALITY OF REGRESSIVE EXPECTATIONS II
(Independent variable: $\bar{s}(t)-s(t)$; Long-run equilibrium *PPP*)

OLS Regressions of Survey Prediction Errors: $s^e(t+1)-s(t+1)=a+\theta(\bar{s}(t)-\bar{s}(t))$

Data Set	Dates	θ	*DF*	t: $\theta=0$	R^2	*F* test $a=0, \theta=0$
Economist 3-Month	6/81–12/85	−0.2041 (0.1100)	184	−1.86[a]	0.28	1.24
MMS 3-Month	1/83–10/84	−0.0335 (0.1387)	182	−0.24	0.01	6.01[c]
Economist 6-Month	6/81–12/85	−0.4344 (0.2252)	174	−1.93[a]	0.29	1.49
Amex 6-Month	1/76–8/85	0.0343 (0.1643)	45	0.21	0.00	1.78
Economist 12-Month	6/81–12/85	−0.5090 (0.2227)	149	−2.29[b]	0.37	6.48[c]
Amex 12-Month	1/76–8/85	0.4278 (0.2412)	40	1.77[a]	0.26	0.85

Notes and footnotes: See Table 8.

expectation: these expectations appear to be overly adaptive.[20]

In Tables 6 and 7 we found that investors expected the spot rate to regress over the subsequent year toward a long-run equilibrium, at a rate of up to 24 percent of the existing gap. In Tables 10 and 11 we test whether this regressive expectation is borne out by reality. An earlier version of this paper that included data only up to March 1985 showed that the *Economist* data were overly regressive. But now in both the *Economist* and MMS data the actual spot rate on average regressed toward equilibrium to an even greater extent than investors expected. In the case of the *Economist* twelve-month data, the highly significant coefficient is evidence that investors systematically un-

[20] Stephen Marris (1985, pp. 120–22) uses the *Economist* survey data and argues that expectations are overly adaptive in that a forecasting strategy of putting less weight on the contemporaneous spot rate would ultimately be vindicated in the long run.

derestimated the degree of regressivity. But the results are dominated by the peaking of the dollar in 1985. When the years 1976–78 are included (the Amex sample), there is on average no tendency for the spot rate to regress toward equilibrium. Again, the finding of systematic expectational errors is fairly robust, but the sign is sensitive to the precise sample period.

V. Thoughts on "The" Expected Exchange Rate

Several considerations suggest that, if we were to reject the hypothesis of rational expectations, the alternative hypothesis would have to be more complex than the simple models considered above. In Table 3, we found that investors systematically overpredicted the depreciation of the dollar in the 1980's, and systematically underpredicted its depreciation in the late 1970's. Similarly, there was a consistent tendency for investors to overestimate the speed of regression before 1985 and to underestimate it thereafter. Such findings suggest the possibility that the nature of the forecasting bias changes over time. Investors could even be rational, and yet make repeated mistakes of the kind detected here, if the true model of the spot process is evolving over time. There is nothing in our results to suggest that it is easy to make money speculating in the foreign exchange markets.

Another puzzle is that the gap between the forward discount and the expected rate of depreciation in the survey data is so large, an average of 7 percent for the *Economist* six-month data. To explain the gap as a risk premium would require (a) that assets denominated in other currencies were perceived in the early 1980's as riskier than assets denominated in dollars, and (b) that investors are highly risk averse. An alternative is the possibility that investors do not base their actions on a single homogeneous expectation such as regressive expectations. If expectations are heterogeneous, then the forward discount that is determined in market equilibrium could be a convex combination of regressive expectations and other forecasts that are closer to static expectations.

There is a third clue that expectations are more complex than a simple homogeneous model, such as those estimated above. In our results, the three-month survey data exhibit a lower speed of regression toward the long-run equilibrium, even when annualized, than do the six-month data, and the six-month survey data exhibit a lower speed of regression than do the twelve-month data. This pattern in the term structure suggests the possibility that those investors who think longer-term tend to be the ones who subscribe to regressive expectations, and those who think shorter-term tend to be the ones who subscribe to forecasts that are closer to static expectations.

In the present paper we have treated exchange rate expectations as homogeneous, for the simple reason that almost all the literature, both theoretical and empirical, does so. Our goal here has been to test standard propositions about "the" expected rate of depreciation, whether it is nonzero, whether it is inelastic, whether it is rational, etc. But in fact, each forecaster has his or her own expectation. The *Economist* six-month survey, for example, reports a high-low range around the median response; it averages 15.2 percent for the five exchange rates.[21] Different models may be in use at one time. We believe that heterogeneous expectations and their role in determining market dynamics are important areas for future research.[22]

VI. Conclusions

To summarize our findings:

1) Exchange rate expectations are not static. The observed nonzero forward discount numbers, far from being attributable to a positive risk premium on the dollar during the recent period, have *understated*

[21] Such heterogeneity across investors can still be compatible with a well-defined market expectation. Mark Rubinstein (1974) gives conditions under which agents with different beliefs may be aggregated to form a composite investor with preferences exhibiting rational expectations.

[22] Possibilities in this line of research are contained in Roman Frydman and Edmund Phelps (1983) and our paper (1986b).

the degree of expected dollar depreciation, which was consistently large and positive.

2) Exchange rate expectations do not exhibit bandwagon effects. We find that the elasticity of the expected future spot rate with respect to the current spot rate is in general significantly less than unity; expectations put positive weight on the "other factor," regardless of whether it is the lagged spot rate (distributed lag expectations), lagged expected rate (adaptive expectations), or the long-run equilibrium rate (regressive expectations). The general finding of inelastic expectations is important because it implies that a current increase in the spot exchange rate itself generates anticipations of a future decrease, as in the overshooting model, which should work to moderate the extent of the original increase. Speculation is stabilizing.

3) While expected depreciation is large in magnitude, the actual spot exchange rate process may be close to a random walk, giving rise to unconditional bias in the survey forecast errors during the 1980s. In view of point 2, a spot process that is close to a random walk would suggest that expectations are less elastic than is rational. Indeed, we find statistically significant bias conditional on, for example, lagged expectational errors. This is the same finding common in tests of efficiency in the forward exchange market, but it now cannot be attributed to a risk premium.

4) The nature of the rejection of rational expectations strongly depends on the sample period. During the 1981–85 period, the actual spot process did not behave according to investors' expectations that the currency would return toward its previous equilibrium, but, after February 1985, the dollar depreciated at a rate in excess of what was expected. It seems likely that the actual spot rate process is more complicated than any of the models tested here.

5) While the present paper adopted the standard theoretical and empirical framework that assumes homogeneous expectations, a number of clues suggest that investigating heterogeneous investor expectations would be a useful avenue for future research.

DATA APPENDIX

Here we describe the construction of the *Economist*, Amex, and MMS data sets more specifically.

The *Economist Financial Review* conducted 38 surveys beginning in June 1981 through December 1985. Surveys took place on a specific day on which the foreign exchange markets were open. Respondents were asked for their expectations of the value of the five major currencies against the dollar in three-, six-, and twelve-months time. We carefully matched a given day's survey results with that day's actual spot and forward rates, and with actual spot rates as close as possible to 90, 180, and 365 days into the future.

The *Amex Bank Review* has conducted 12 surveys beginning in January 1976 through July 1985. Respondents were asked for their expectations of the value of the same five currencies in six- and twelve-months time. The first three surveys, however, included only the pound and the mark. Future foreign exchange market realizations were matched in a manner similar to that used for the *Economist* data. Amex Bank surveys were conducted by mail, and hence it was impossible to pick specific days which were used by all respondents as reference points with any degree of certainty. Since exchange rates vary so much within a month, two methods of choosing the contemporaneous spot rate (and the corresponding future rates respondents were predicting) were employed. First, single days within the survey period were selected. Second, 30-day averages of daily rates were constructed to encompass the entire survey period. Since both methods yielded very similar quantitative results in the body of the paper, the results from the latter Amex data set are reported only in the NBER working paper version.

Between January 1983 and October 1984, MMS conducted 47 surveys (one each two weeks) of the value of the dollar against the pound, mark, Swiss franc, and yen in three-months' time. Matching of actual spot and forward rates was done in a manner similar to that used for the *Economist* survey.

Actual market spot and forward rates were taken from DRI. They represent the average of the morning bid and ask rates from New York. Lagged exchange rates (used for extrapolative expectations) are market rates approximately 90 days before survey dates.

Specific dates on which the surveys were conducted, and for which actual market data was obtained, are contained in Tables A1, A2, and A3 in our paper (1986a).

REFERENCES

Bilson, John, "The Speculative Efficiency Hypothesis," *Journal of Business*, July 1981, *54*, 431–51.

______, "Macroeconomic Stability and Flexible Exchange Rates," *American Economic Review Proceedings*, May 1985, *75*, 62–67.

Branson, William, Halttunen, Hannu and Masson,

Paul, "Exchange Rates in the Short Run: The Dollar-Deutschemark Rate," *European Economic Review*, December 1977, *10*, 303–24.

Cornell, Bradford, "Spot Rates, Forward Rates, and Exchange Market Efficiency," *Journal of Financial Economics*, August 1977, *5*, 55–85.

Dominguez, Kathryn, "Are Foreign Exchange Forecasts Rational: New Evidence from Survey Data," International Finance Discussion Paper Series, No. 241, May 1986, forthcoming *Economic Letters*.

Dooley, Michael and Shafer, Jeffrey, "Analysis of Short-Run Exchange Rate Behavior: March 1973 to November 1981," in D. Bigman and T. Taya, eds., *Exchange Rate and Trade Instability: Causes, Consequences, and Remedies*, Washington: International Monetary Fund, 1983.

Dornbusch, Rudiger, (1976a) "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy," *Scandinavian Journal of Economics*, May 1976, *78*, 255–79.

______, (1976b) "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–76.

Engel, Charles M., "Testing for the Absence of Expected Real Profits from Forward Market Speculation," *Journal of International Economics*, November 1984, *17*, 299–308.

Evans, George W., "A Test for Speculative Bubbles and the Sterling-Dollar Exchange Rate: 1981–84," *American Economic Review*, September 1986, *76*, 621–36.

Frankel, Jeffrey, "Tests of Rational Expectations in the Forward Exchange Market," *Southern Economic Journal*, April 1980, *46*, 1083–101.

______, "The Dazzling Dollar," *Brookings Papers on Economic Activity*, 1:1985, 199–217.

______ **and Froot, Kenneth A.,** (1986a) "Using Survey Data to Test Some Standard Propositions Regarding Exchange Rate Expectations," Research Papers in Economics No. 86–11, Institute of Business and Economic Research, University of California-Berkeley, April 1986.

______ **and** ______, (1986b) "The Dollar as a Speculative Bubble: A Tale of Chartists and Fundamentalists," NBER Working Paper, No. 1854, March 1986.

Friedman, Milton, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953, 157–203.

Froot, Kenneth A. and Frankel, Jeffrey A., "Interpreting Tests of Forward Discount Unbiasedness Using Survey Data on Exchange Rate Expectations," NBER Working Paper No. 1963, July 1986.

Frydman, Roman and Phelps, Edmund S., *Individual Forecasting and Aggregate Outcomes: "Rational Expectations" Examined*, New York: Cambridge University Press, 1983.

Hansen, Lars and Hodrick, Robert, "Forward Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, October 1980, *88*, 829–53.

Huang, Roger, "Some Alternative Tests of Forward Exchange Rates as Predictors of Future Spot Rates," *Journal of International Money and Finance*, August 1984, *3*, 157–67.

Kouri, Pentti J. K., "The Exchange Rate and the Balance of Payments in the Short Run and the Long Run: A Monetary Approach," *Scandinavian Journal of Economics*, No. 2, 1976, *10*, 280–304.

Levich, Richard, "Analyzing the Accuracy of Foreign Exchange Advisory Services: Theory and Evidence," in R. Levich and C. Wihlborg, eds., *Exchange Risk and Exposure*, Lexington: D. C. Heath, 1979.

Lovell, Michael C., "Tests of the Rational Expectations Hypothesis," *American Economic Review*, March 1986, *76*, 110–24.

Marris, Stephen, *Deficits and the Dollar: The World Economy at Risk*, Washington: Institute for International Economics, December 1985.

Meese, Richard and Rogoff, Kenneth, "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?," *Journal of International Economics*, February 1983, *14*, 3–24.

Mincer, Jacob, "Models of Adaptive Forecasting," in his *Economic Forecasts and Expectations*, New York: Columbia University Press, 1969.

Newey, Whitney and West, Kenneth, "A Simple, Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," Woodrow Wilson School Discussion Paper No. 92, May 1985.

Nurkse, Ragner, *International Currency Experience*, Geneva: League of Nations, 1944.

Parks, R. W., "Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated," *Journal of the American Statistical Association*, 1967, *62*, 500–09.

Rubinstein, Mark, "An Aggregation Theorem for Securities Markets," *Journal of Financial Economics*, September 1974, *1*, 225–44.

Stockman, Alan C., "Risk, Information and Forward Exchange Rates," in J. A. Frenkel and H. G. Johnson, eds., *The Economics of Exchange Rates*, Reading: Addison Wesley, 1978.

American Express International Banking Corporation, *Amex Bank Review*, London, various issues, 1976–85.

The Economist Newspaper, Ltd., *Financial Report*, London, various issues, 1981–86.

Money Market Services, Inc., *Currency Market Services*, various issues, 1983–86.

The Welfare Effects of Third-Degree Price Discrimination in Intermediate Good Markets

By MICHAEL L. KATZ*

This paper examines third-degree price discrimination by an intermediate good monopolist selling to downstream firms that differ in their abilities to integrate backward into supply of the input. It is shown that discrimination may lead to all buyers facing higher prices, and conditions under which discrimination reduces welfare by lowering total output are presented. It is shown that discrimination may raise welfare in some cases by preventing socially inefficient backward integration.

"I come to exhume Robinson-Patman, not to praise it."

American public policy towards price discrimination is largely concerned with discrimination in intermediate goods markets. In fact, the principal legislation covering price discrimination, the Robinson-Patman Act, was introduced with the explicit intention of protecting small businesses from "unfair" advantages possessed by large buyers in intermediate goods markets. Throughout the life of the Robinson-Patman Act, there has been considerable debate among both economists and those responsible for enforcing the Act about whether intermediate good price discrimination has beneficial or adverse effects. In recent years, a consensus has developed that the effects of intermediate good price discrimination are beneficial and thus such discrimination should not be proscribed. Consequently, the number of Robinson-Patman cases brought by the Federal Trade Commission fell markedly in the 1970's.[1]

The prevailing view is the following one. Price discrimination allows the upstream supplier to make selective price cuts to those customers with respect to which it has the least market power, and these cuts would not be profitable if price discrimination were not allowed (i.e., if all prices had to be lowered). Hence, under price discrimination, an intermediate good monopolist will lower prices to some buyers without raising prices to the remaining buyers. As a result, price discrimination leads to lower final good prices from which the ultimate consumers benefit.

This line of argument should not be a convincing one. Its mirror image can be made to attack price discrimination: when allowed to price discriminate, the upstream supplier makes selective price increases to those customers with respect to which it has the greatest market power, where these increases would not have been profitable if price discrimination were forbidden. In this view, price discrimination leads to higher prices and lower welfare.[2]

*Department of Economics, Princeton University, Princeton, NJ 08544. I thank Charles Berry, John Panzar, Richard Quandt, Carl Shapiro, Robert Willig, and seminar participants at Northwestern University, University of Pennsylvania and Princeton University for helpful suggestions and comments. This material is based on work supported by the National Science Foundation under grants no. SES-8209412 and SES-8510746.

[1] See Richard Posner (1976, Table 1, p. 32).

[2] Some claim that discrimination induces cartel instability by making it more profitable for member firms to cheat. It also has been argued that price discrimination encourages entry; a firm that is established in one market can enter a new market, which may necessitate setting a low price, without having to lower the price in its home market. The U.S. Department of Justice (1977), among others, presents variants of these arguments. But it can be argued that cartels are strengthened because firms can compete where cartelization is bound to fail while maintaining collusive prices to those consumers where cartelization is viable. Similarly, price discrimination may deter entry because a multimarket incumbent can credibly threaten to respond more aggressively to a single-market entrant when the incumbent can make targeted price cuts directed against the entrant.

Despite the long and extensive history of the debate, there has been no formal modeling of third-degree price discrimination in intermediate good markets (i.e., markets where the seller can set buyer-specific uniform prices).[3] Ironically, there have been many welfare analyses of one market structure to which current U.S. antidiscrimination legislation does *not* apply—third-degree price discrimination by a monopolist in a *final* good market.[4] Two excellent analyses of final good monopoly have been conducted by Richard Schmalensee (1981b) and Hal Varian (1985). The basic findings of these papers are that: (a) when there are two classes of buyers, discrimination leads to a higher price for one class and a lower price for the other; and (b) profit-maximizing third-degree price discrimination may raise or lower welfare in comparison with profit-maximizing uniform pricing, where the sign of the welfare effect depends on the curvature of the demand curves.

One might suspect that these results carry over to intermediate good markets, and thus that the truth about the welfare effects of intermediate good price discrimination lies somewhere between the two extreme views presented above. I show below, however, that models of final good markets are inappropriate for the analysis of intermediate good markets—the effects of intermediate good price discrimination may fall outside of the two extremes already sketched.[5]

There are two fundamental differences between final and intermediate good markets. First, in an intermediate good market, unlike a typical final good market, the buyers' demands for the product are interdependent. The profits of any given downstream firm and its demand for an input are functions both of the price that the firm pays for the input and of the prices that the buyer's product-market rivals pay. A second important difference between intermediate good markets and final good markets is that buyers of inputs often have the ability to integrate backward into supply of the good.

In this paper, I analyze the effects of third-degree price discrimination in an analytical framework that explicitly recognizes both of these differences. The intermediate good is an input into the production of a homogeneous final good that is sold downstream in a set of independent local markets. There are two sellers in each of the downstream markets. One store in each market is a branch of a chain that sells the final good in all of the downstream markets, while the other seller is a local store that operates solely in that market. This structure may be thought of as a stylization of the competition between chain stores and local stores that was the impetus for passage of the Robinson-Patman Act. Because of its larger demand for the input and the economies of scale in production of the intermediate good, the chain has a stronger threat of backward integration than do the local stores. This difference in the buyers' ability to integrate backward profitably gives rise to the upstream seller's incentives to discriminate.

Two types of welfare effect may occur in the model considered here. First, the total amount of the intermediate good sold, and thus the total amount of the final good consumed, may vary across the discriminatory and nondiscriminatory pricing regimes. Second, the costs of production for a given level of output may be higher under one pricing regime than the other, due to differences in the chain's integration decision. When the chain integrates under either regime, the two outcomes are identical; the upstream incumbent sells only to the local stores and has no need to engage in price discrimination. Thus, in this case neither welfare effect arises.

[3] Janusz Ordover and John Panzar (1980, 1982), Panzar and David Sibley (1983), and Richard Schmalensee (1981a) have examined the related case of an intermediate good monopolist who charges a nonuniform price schedule to a set of buyers who have interdependent demands. These authors do not consider discrimination motivated by buyers' threats of backward integration.

[4] The Robinson-Patman Act concerns harm to competition. But in the case of a final good monopoly, there is no competition among either sellers or buyers of the good.

[5] Similarly, in their work on nonuniform pricing, Ordover and Panzar found that intermediate good markets may be strikingly different than final good markets.

When the chain does not integrate under either regime, only the first type of welfare effect arises. Here, one sees the most striking example of the difference between intermediate and final good markets. Under reasonable conditions, intermediate good price discrimination leads to higher input prices being charged to *all* buyers, a result that never would arise in a corresponding model of a final good market. When both prices rise, output and welfare fall. In other cases, price discrimination leads to one price falling and the other rising, as it would if one were considering a final good market. Surprisingly, the price charged to the chain store may be the one that rises. The main welfare finding is that the adverse effects of the price increase outweigh the beneficial effects of the price decrease when the downstream firms are Cournot competitors; again, price discrimination lowers output and welfare. Unlike final good markets, this result is due to the nature of competitive interaction in the downstream markets, not the curvature of the demand functions.

The final case to consider is that of a market in which one pricing regime leads to chain integration while the other does not. Here, the second type of welfare effect arises. I show that if integration occurs in only one regime, then it must be the case that integration occurs when price discrimination is banned, but not when it is allowed. Consequently, a potential benefit of price discrimination is that it prevents socially inefficient integration, and there are markets in which price discrimination raises welfare.

The remainder of this paper is organized as follows. The formal analytical framework is presented in Section I. The chain's integration decision is discussed in Section II. The seller's choice of input prices is covered in Section III, and Section IV contains the main welfare results. The more formal aspects of the analysis are relegated to the Appendix which also generalizes the results presented in the text.

I. The Analytical Framework

Legal cases involving price discrimination in the sale of intermediate goods typically concern pricing schemes under which customers with large individual purchases or large cumulative purchase volumes receive lower prices than do customers making small purchases. In order to understand the welfare effects of intermediate good price discrimination, it is essential to understand why it is the case that large buyers often are charged lower prices than are small buyers.[6]

In the model analyzed below, a large buyer receives lower prices because there are economies of scale in finding an *alternative* source of supply, and thus the threat of finding an alternative is stronger when it is made by a high-volume buyer. In particular, a buyer may threaten to engage in self-supply using a production technology that exhibits economies of scale. Alternatively, one could consider a model in which a buyer would threaten to go to another upstream supplier. In this case, the economies of scale on the buyers' side of the market could arise from a buyer having to incur fixed costs to find and contract with another supplier. Or, the buyer might have to bear fixed costs to modify its production line to utilize an alternative variant of the input.

One could allow for customer-specific or transaction-specific economies of scale on the seller's side of the market (for example, fixed costs of selling effort, contracting, or delivery). When there are economies of scale on the sellers' side of the market, lower prices for high-volume buyers may merely reflect the underlying cost differences and need not be evidence of price discrimination. In fact, the high-volume buyer may be discriminated against while receiving a lower price if that price does not reflect all of the cost savings that result from the transactional economies of scale. While supply-side

[6] A typical story told to explain quantity discounts is that losing a high-volume customer is more costly to the seller than is losing a low-volume customer. This argument does not, however, answer the question of where the customer goes if it leaves its current supplier. To make a credible threat to leave, the buyer must have an alternative source of supply. A threat made by a large buyer will be stronger than that of a low-volume buyer only if there are transactional economies of scale in either buying or selling the input. This point is discussed further in my working paper (1985).

economies of scale can explain the existence of price differentials, they are not sufficient to explain why the markup of price over marginal or incremental cost should vary in favor of large buyers. Since my interest here is in price discrimination, not price differentials, I assume that there are no supply-side economies of scale in transactions.

Turning to the formal model, there is an upstream industry whose output is an input into the production of good X by a downstream industry. Initially, the output is produced by the single incumbent upstream producer at a constant, positive marginal cost of c. One could allow for fixed costs of production, but as long as they are sunk they will have no effect on the upstream supplier's behavior.

Each downstream producer requires a fixed amount of the input to produce a unit of the final good. In order to focus on differences in the strengths of buyers' integration threats as the motivation for discrimination, I assume that all of the downstream firms are equally efficient producers. Given these assumptions, there is no further loss in generality by assuming that there are no other inputs to production and defining the units of the goods so that one unit of the input is needed to produce one unit of the final good.

The downstream industry sells its output in $K \geq 2$ identical, independent markets. The products of the firms in a given market are perfect substitutes, and the market price is given by the inverse demand curve $P[\cdot]$, which is a function of the total output sold in *that market*. For expositional convenience, in the text I assume that $P[\cdot]$ is linear. This assumption is dropped in the Appendix.

There are two producers in each downstream market. Firm 1, "the chain," is active in all K downstream markets. The other firm in any given market, "the local store," operates in that market only. Because all of the markets are identical, I examine the behavior of firm 1 and its competitor, firm 2, in a single, representative market.

The firms are Cournot competitors; each downstream firm chooses its output under the assumption that its rival's output is fixed. Let $x_1[m_1, m_2] = x[m_1, m_2]$ denote the equilibrium output of firm 1 when it has a marginal input cost m_1 and firm 2 faces an input cost of m_2.[7] Given homogeneous demand and Cournot behavior, the output of firm 2 is given by $x_2[m_1, m_2] = x[m_2, m_1]$. Let X denote the equilibrium level of total output in a single market; $X[m_1, m_2] = x[m_1, m_2] + x[m_2, m_1]$. A straightforward comparative statics exercise demonstrates that an increase in m_i lowers X if firm i initially had positive output.

The downstream profits earned in a single market by firm i are

$$(1) \quad \pi[m_i, m_j] = x[m_i, m_j] \times \{P[X(m_i, m_j)] - m_i\}.$$

It is trivial to verify that for all input prices such that $x[m_i, m_j] > 0$: (a) $\partial\pi[m_i, m_j]/\partial m_i < 0$; and (b) $\partial\pi[m_i, m_j]/\partial m_j \geq 0$, with strict equality only if $x[m_j, m_i] = 0$.

If firm i purchases the input from the upstream producer, m_i is equal to w_i, the price that the upstream producer charges firm i for the input, and the firm's profits per market are $\pi[w_i, m_j]$.

The downstream firms may integrate backward and produce the input themselves. The same upstream production technology is available to all of the downstream firms. The key feature of the technology is that it exhibits increasing returns to scale. For simplicity, I assume that the cost of producing y units of the intermediate good is $F + vy$, where F and v are positive constants with $v \geq c$. The cost F is a fixed cost that must be sunk for production to take place. If firm i engages in backward integration, then $m_i = v$, the marginal cost of self-supply. An integrated firm earns an average of $\pi[v, m_j] - F/k_i$ plus any profits from selling the input to other downstream firms, where k_i is the

[7] If $P[X] = \alpha - X$, with α a positive constant greater than c, then $x[m_i, m_j] = \max\{(\alpha - 2m_i + m_j)/3, 0\}$. I use the $x[m_i, m_j]$ notation throughout the text to emphasize that this analysis holds more generally than just the linear demand case.

number of downstream markets in which firm i is present.[8]

It is assumed that no downstream firm has integrated backward yet (i.e., the fixed cost F has not been sunk). Given the assumption that $v \geq c$, it is not socially efficient for any downstream firm to integrate; integration will lead to higher industrywide costs of producing a given level of total output.

II. The Integration Decision

Given the economies of scale in production of the input, the downstream firms typically differ in the strength of their integration incentives. Suppose, first, that a firm that integrates is prevented (either by the courts or by transaction costs) from selling the intermediate good to other downstream producers.[9] It is easy to see that when sales to other downstream firms are blocked, the chain is likely to have a more severe threat of integration than does a single-market firm—the chain can take greater advantage of the economies of scale in self-supply $(F/K < F)$.

Even in markets where the integrator can serve other downstream firms, and thus a local store has an equally large potential market for its upstream output, the chain's integration incentives are higher. When an integrated firm sells to other downstream firms, some of the benefits of integration accrue to these buyers. When "selling" the input to itself, however, an integrated firm appropriates all of the benefits of integration. Put another way, integration leads to upstream competition which confers a positive pecuniary externality on downstream firms. Since it is active in more downstream markets, the chain is better able to internalize this externality.

In the light of the relationship between the incentives of the local store and the chain store, I simplify the exposition by assuming that the local store finds integration unprofitable.[10] In deciding whether to integrate backward, the chain correctly recognizes that no other firm will enter the upstream market.

The chain makes its integration decision by comparing its expected profits with and without integration. While many factors may enter into the chain's predictions of these profit levels, the current input prices are the only factors that the upstream supplier can manipulate at the time that the integration decision is made. Hence, it is useful to express expected integration profits as a function solely of current prices: $\pi^e[w_1, w_2]$. These expected integration profits may include both those profits earned from sales of the final good and those earned from sales of the input to other downstream firms.

The chain is limited to making credible threats of integration. That is, the chain integrates if and only if doing so raises its expected profits:

$$(2)\quad \{\pi^e[w_1, w_2] - F/K\} - \pi[w_1, w_2] \geq 0.$$

For any w_2, let $I[w_2]$ denote the value of w_1 such that the chain is indifferent between integration and nonintegration when the current prices are $I[w_2]$ and w_2, that is, this price pair satisfies equation (2) with equality.

A key characteristic of the market is whether the integration frontier is upward or downward sloping. The slope of the integration frontier depends on whether changes in w_1 and w_2 raise or lower the left-hand side of equation (2).

First, consider the effects that the upstream supplier's raising the chain's input price has on the chain's integration incentives. If it integrates, the chain faces a marginal input cost of v. Thus, the only effects that w_1 can have on expected integration profits must come through signalling effects that w_1 may have on the expected post-

[8] When $v > c$, a downstream firm might sink F and then continue to purchase the input from the upstream incumbent at a price of v.

[9] For example, a U.S. District Court found that it was illegal for A&P to sell to competing grocery stores (Morris Adelman, 1953). Alternatively, the chain may believe that the upstream incumbent would undercut it on any sales to the local stores, and thus expect zero profits from sales to its product-market rivals.

[10] In the Appendix, I treat the case in which both types of firm have a credible threat of integration.

integration level of w_2. For example, the chain may take a high value of w_1 as a signal that the upstream supplier has high production costs and thus would set the local store price at a high level if the chain were to integrate backward. To the extent that expected integration profits are affected by a rise in w_1, they are likely to increase.[11] Clearly, raising the price charged to the chain lowers its nonintegration profits. Therefore, raising w_1 increases the chain's incentives to integrate backward.

It is much less clear whether raising the local store's price raises or lowers the chain's incentivè to integrate. Depending on the market institutions, modes of upstream competition after integration, and the set of feasible contracts, either case may hold.

Suppose that the upstream incumbent can adjust its prices instantaneously in response to integration and that the chain has complete information about the market conditions (for example, the upstream supplier's costs, the demand for the input, and the aggressiveness of the upstream supplier's management) that will prevail after integration. In this case, the value of w_2 has no effect on the chain's prediction of the post-integration input prices and profits. But raising w_2 increases the chain's nonintegration profits, $\pi[w_1, w_2]$. Hence, raising w_2 reduces the chain's integration incentives, and $I[\cdot]$ is upward sloping, as Figure 1 illustrates.

Now, suppose that the chain has poor information about the post-integration market conditions (for example, it does not know the upstream supplier's costs). In this case, current prices may serve as signals. Through signalling effects, an increase in the local store input price may raise the chain's integration incentives by raising its expected integration profits by more than it raises the chain's nonintegration profits.

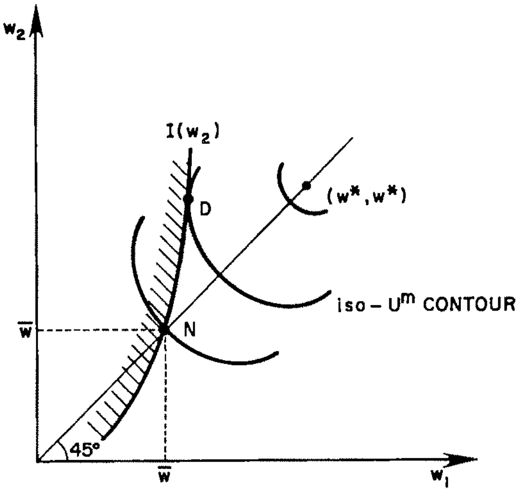


FIGURE 1

Similarly, if the upstream supplier's prices cannot be adjusted in response to integration, then a high current price may raise the chain's integration incentives. There are two reasons for this relationship. First, under integration, the chain is more profitably able to undercut the upstream incumbent when the latter has its price fixed at a high level. Second, by setting w_2 at a higher level, the upstream incumbent induces the local store to reduce its output. This contraction raises the final good's price, increasing the chain's profits. The increase in the chain's profits due to the rise in price is greater when it integrates than when it does not, because under integration the chain has a low marginal input cost and thus produces a high level of output. Therefore, an increase in the local store's price makes integration relatively more desirable to the chain.

When an increase in w_2 raises the chain's integration incentives, the upstream incumbent has to lower w_1 to forestall integration by the chain. In this case, $I[\cdot]$ is downward sloping, as illustrated in Figure 2.

The case of $I[\cdot]$ upward sloping, which arises when the integration rule takes the form "integrate if and only if $\pi[w_1, w_2]$ is less than some trigger level," seems to me to

[11] This is a variant of the modern theory of limit pricing. Paul Milgrom and John Roberts (1982) analyze a market where an informational asymmetry between an incumbent and a potential entrant gives rise to similar signalling. One can construct examples where high initial prices would signal that entry is unprofitable. I will not consider such cases here, other than to note that all of the propositions stated in the paper hold for these cases as well.

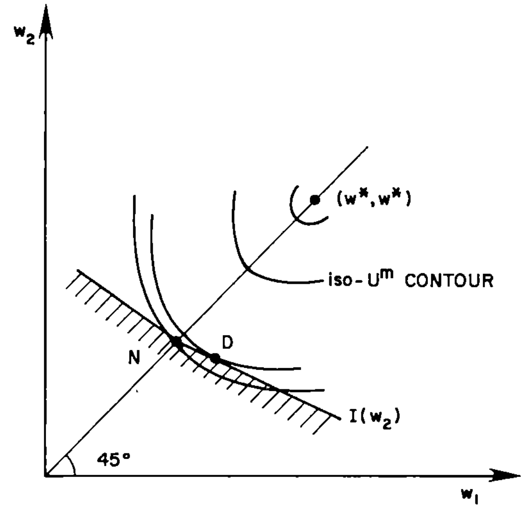


FIGURE 2

be the more natural of the two extremes. Actual markets probably fall somewhere between these two poles, however; prices can be adjusted, but only slowly, and there is a role for signalling or limit pricing. Hence, the analysis below considers both cases in which $I[\cdot]$ is upward sloping and cases in which the integration frontier is downward sloping.

III. The Upstream Supplier's Choice of Prices

The seller chooses w_1 and w_2 to maximize its profits, taking the chain's integration rule into consideration. When the upstream supplier can engage in price discrimination, the maximal profits that the supplier can earn without inducing integration are given by the solution to

$$(3)\quad \max_{w_1,w_2} U^m[w_1,w_2]$$

$$\equiv (w_1-c)x[w_1,w_2]+(w_2-c)x[w_2,w_1]$$

subject to

$$\pi[w_1,w_2] \geq \pi^e[w_1,w_2]-F/K.$$

When it has an upstream monopoly m, the supplier's profit function $U^m[\cdot,\cdot]$ is symmetric. Given that the downstream firms are Cournot competitors facing a linear market demand curve, the seller's profit function is quasi concave over the set of prices for which both downstream firms have positive output levels.

Figure 1 illustrates the solution to this problem for the case of $I[\cdot]$ upward sloping. By the symmetry and quasi concavity of $U^m[\cdot,\cdot]$, the supplier would like to charge w^* to both downstream firms. But this pair of prices would induce the chain to integrate (for the remainder of the analysis, I will assume that the threat of integration makes (w^*,w^*) infeasible). In order to prevent firm 1's integrating, the supplier must choose a set of prices in the shaded region. The supplier's profits are maximized by choosing w_1 unequal to w_2 at point D.

If the upstream supplier sets input prices that induce the chain to integrate, it earns profits given by the solution to

$$(4)\quad \max_{w_1,w_2} U^d[w_1,w_2],$$

where $U^d[w_1,w_2]$ denotes the upstream incumbent's profits given that it originally announced prices (w_1,w_2) and chain integration takes place (i.e., there is an upstream duopoly d). Making the natural assumption that, for any $(w_1,w_2), U^d[w_1,w_2] \leq U^m[w_1,w_2]$, if the solution to equation (4) yields profits that are greater than those under the solution to equation (3), it must be the case that the input prices that maximize equation (4) were not feasible under equation (3). That is, these prices will, in fact, induce chain integration.

When price discrimination is forbidden, the upstream producer must offer the input to both downstream firms at the same price. Let w_b denote the common price charged to the downstream firms when discrimination is not allowed. If the upstream producer chooses not to induce integration, then both downstream firms purchase the input. In this case, the supplier sets w_b at as high a level as satisfies the integration constraint: $w_b = I[w_b]$. In terms of Figure 1, when price discrimination is infeasible, the supplier must choose a price pair that is on the 45° ray coming from the origin. The highest feasible level of w_b is illustrated by point N, $(\bar{w}, \bar{w})$.

Alternatively, the upstream firm may elect to set a price such that $I[w_b] < w_b$ and the chain integrates. In this case, the upstream incumbent chooses w_b to maximize $U^d[w_b, w_b]$. The supplier may choose such a price in order to exploit its market power with respect to the local stores. If it were to prevent chain integration, the upstream incumbent would have to set w_b at a low level. This intuition is most clear in markets where $I[\cdot]$ is downward sloping. In such markets, if prices (w°, w°) induce integration, but $(\bar{w}, \bar{w})$ do not, then it must be the case that $w^{\circ} > \bar{w}$. Thus, in a market where the upstream supplier sets w_b at a level that induces chain integration, the price must be higher than the one at which upstream profits would be maximized subject to the constraint that chain integration not be induced.

IV. Welfare Analysis

Given the assumptions of my model, there are two types of welfare effects to which price discrimination may give rise: 1) the total amount of the intermediate good sold, and thus the total amount of the final good consumed, may vary across the discrimination and no-discrimination pricing regimes; and 2) the chain's integration decision, and thus the costs of producing a given level of total output, may differ in the two regimes.

Whether these two types of welfare effect are present in any given comparison of regimes depends on whether the upstream incumbent induces integration by the chain. If the upstream incumbent finds it optimal to induce integration both with and without price discrimination, then the two outcomes are identical since the upstream incumbent serves only one type of buyer (local stores) in each case. Thus, neither welfare effect arises.

A. *Integration in Neither Regime*

In markets where it is not profitable for the upstream incumbent to induce chain integration under either pricing regime, only the first type of welfare effect arises. The following key technical result is proved in the Appendix:

LEMMA 1: *Suppose there is no integration under either pricing regime. If both firms are active under price discrimination, then $(w_1 + w_2)/2 > w_b$. If only firm j is active under discrimination, then $w_j > w_b$.*

By the quasi concavity of $U^m[\cdot, \cdot]$, in either regime the upstream supplier chooses input prices that are on the chain's integration frontier. For markets in which raising w_2 lowers the chain's integration incentives, the positive slope of the integration frontier implies that either both prices rise or both fall (see Figure 1). By Lemma 1, both prices rise.

The intuition behind this result is the following. The upstream supplier must set input prices that yield the chain sufficient profits to make integration undesirable. At the no-discrimination equilibrium (point N in Figure 1), the upstream seller would like to increase both prices, but it cannot raise the prices equally without inducing integration by the chain. And, absent the ability to discriminate, the seller must move the two prices together. When discrimination is allowed, however, the seller can raise the price charged to the local store alone. This price increase raises the profits of the seller. Moreover, by inducing the local store to contract its output, the price increase raises the profits of the chain. Thus, the seller can raise the price charged to the chain (albeit by a smaller amount) without inducing integration. Both prices rise (to point D). Clearly, total output falls and the price rises in the final good market.

When an increase in the local store input price increases the chain's integration incentives, $I[\cdot]$ is downward sloping and price discrimination raises one input price while lowering the other. Figure 2 illustrates this case. It is not immediately obvious which price change has the dominant effect on the level of output. Augustin Cournot (1897) has shown that, if the downstream firms are Cournot competitors, then the level of total output can be expressed as a decreasing function of the average price charged to the firms that are active producers in equilibrium. Hence, by Lemma 1, total output is lower at the discriminatory prices than at the nondiscriminatory prices. On balance, the adverse effect of the rising input price dominates.

Price discrimination lowers total output whether the chain's integration frontier is upward or downward sloping. What happens to welfare? Absent integration, the costs of production (the sum of the costs for the intermediate and final goods) depend only on the total output level, not on the distribution of this output between the two downstream producers. Similarly, given that the good is homogeneous, gross consumer benefits depend only on the total amount of the final good that is sold. Using the fact that the downstream equilibrium price is at least as great as marginal cost, it follows that total surplus falls as total output falls.

Summarizing this analysis:

PROPOSITION 1: *If there is no integration under either regime, then total output and welfare are lower when price discrimination is practiced than when it is forbidden.*

A more general version of this result is stated and proved in the Appendix.

It is instructive to examine the distributional consequences of price discrimination in addition to its aggregate welfare effects. Consumers are harmed by price discrimination since the final good price rises. The upstream incumbent clearly benefits since it has the option of setting w_1 equal to w_2 when price discrimination is allowed. The effects of discrimination on the two downstream producers are less clear and depend on the nature of the integration frontier.

When the frontier is upward sloping, both prices are greater with discrimination than without, and they must fall on a portion of the chain's integration frontier that is above the 45° line in Figure 1. Hence, the chain store pays a lower price than does the local store under price discrimination. The intuition is that the seller would like to set $w_1 = w_2$. The only reason to deviate from this policy is to satisfy the constraint that the chain's profits absent integration are sufficiently high to make integration undesirable to the chain. This constraint may be satisfied either by raising w_2 or by lowering w_1. From $w_1 < w_2$, it follows that the chain store earns greater profits under price discrimination than does the local store.

How do these profit levels compare with their no-discrimination values given that both input prices rise? If the chain's expected profits under integration are independent of the current price levels, then the chain store is not made better off by the fact that the seller can discriminate in its favor—the upstream seller drives the chain's nonintegration profits down to the level of expected integration profits with or without discrimination. The local store is worse off, however, since absent discrimination it earns profits equal to those of the chain ($w_1 = w_2$), but under discrimination the local store earns profits that are lower than those of the chain ($w_1 < w_2$).

When the integration frontier is downward sloping, discrimination raises the input price faced by one producer and lowers the input priced faced by the other one. One downstream firm gains while the other loses. Curiously, as Figure 2 illustrates, the chain store may be the loser when lowering the current value of w_2 greatly lowers the chain's expectations of the post-integration level of the price offered by the upstream incumbent to the local store. This fall in the expected price reduces the chain's expected profits from selling to local stores under integration. Moreover, the chain will expect the local store to have a higher level of output under integration. This increase in the local store's output has the effect of reducing the level of final output that the chain would produce under integration. Thus, the chain would be less able to enjoy the economies of scale in supplying itself. If these effects are sufficiently strong, the upstream incumbent will lower w_2 to make integration undesirable to the chain and then increase w_1 to raise upstream profits.

Opponents of price discrimination warn that discrimination may lead to one store's profits falling by so much that it is driven from the downstream market, allowing the other store to raise its price and exercise market power. The next result states that, in the model considered here, these fears are unfounded; such "predation" does not occur.

PROPOSITION 2: *Suppose that the chain's expected integration profits do not fall as the upstream supplier raises either the price charged to the chain store or the price charged to the local store. Then the upstream incumbent will not create a downstream monopoly by setting prices that drive one of the downstream firms out of business.*

PROOF:

It suffices to show that the upstream supplier would not find it profitable to make the chain a downstream monopolist since it would be more difficult (i.e., less profitable) for the upstream supplier to drive the chain firm out of business given its self-supply opportunities. There are two cases to consider.

(*i*) $I[\cdot]$ *downward sloping*. Suppose that $w_1 < w_2$, $w_1 \leq I[w_2]$ (i.e., there is no integration), and $x_2 = 0$. Then

$$(5)\quad U^m[w_1, w_2] = (w_1 - c)x[w_1, w_2] = (w_1 - c)X[w_1, w_2].$$

Given $I[\cdot]$ downward sloping, lowering the local store's input price to w_1 does not induce chain integration, but does raise unit sales ($\partial X/\partial w_2 < 0$). Since the price at which sales are made, w_1, does not fall, upstream profits rise:[12]

$$(6)\quad U^m[w_1, w_1] = (w_1 - c)X[w_1, w_1] > U^m[w_1, w_2].$$

(*ii*) $I[\cdot]$ *upward sloping*. In this case, a less intuitive proof is needed. (This proof is presented in the Appendix.)

The assumption made in the statement of Proposition 2 is the natural one to make. If there are no signalling effects or price rigidities, then current prices do not affect expected integration profits. If there are signalling effects or price rigidities, then an increase in current prices raises the chain's expected integration profits.

B. *Integration in One Regime Only*

Now, consider the mixed case, where one pricing regime leads to chain integration but the other does not. In deciding whether to induce integration, the upstream supplier compares the maximal value of its profits at prices that prevent integration with its profits when integration takes place. Absent integration, the upstream supplier's profits are greater when price discrimination is allowed than when it is banned. Profits under integration, however, are independent of whether or not discrimination is allowed because the upstream supplier faces only one type of buyer, the local stores. It follows that

PROPOSITION 3: *If the upstream incumbent induces chain integration under only one of the two pricing regimes, integration must be induced in the no-discrimination regime. Hence, in some markets, price discrimination prevents integration and the construction of inefficient upstream production facilities.*

Of course, as before, discrimination gives rise to total output effects. Now, however, the comparison of output levels is more difficult to make because one regime entails upstream duopoly while the other entails upstream monopoly. The comparison of regimes is further complicated by the fact that production efficiency effects arise which may go in the opposite direction of the total output effects. Even if price discrimination leads to lower total output and consumers' surplus in the final good market, this adverse effect may be outweighed by the savings from the avoidance of integration costs. In fact, numerical examples can be constructed to show that if the no-discrimination equilibrium entails integration and the discrimination equilibrium does not, then

[12] The interested reader may note that neither part of the proof relies on the linearity of downstream demand. In fact, the proof of part (*i*) relies only on the properties that lowering w_i raises total sales of the input and does not induce integration for $i = 1$ or 2.

welfare may be greater or less under the no-discrimination regime than under the discrimination regime.

C. *Extensions*

To this point, I have assumed away two potential effects of intermediate good price discrimination. First, I have not allowed for welfare effects arising from changes in the distribution of a given level total output among the downstream firms. For a given level of aggregate output in the downstream market, the distribution of output across firms will vary as the downstream firms' costs (input prices) are shifted by discrimination. Such shifts will have welfare consequences if there are differences in the downstream producers' real costs of production (i.e., their costs of production measured in terms of the real resources used for inputs), or if the goods are heterogeneous. For example, in the case where the chain is the lower-cost downstream firm (that may be why it's the chain in the first place), price discrimination that leads to $w_1 < w_2$ will shift production from the inefficient local store to the efficient chain, raising total surplus for any given level of total output. This positive distribution effect will (at least partially) offset the negative total output effect.

It is interesting to consider the effects of such cost differences in a model in which neither downstream firm has a credible threat of integration.

PROPOSITION 4: *Suppose that neither downstream firm can integrate backward and firm i's marginal cost of production is given by $w_i + d_i$, where d_i is the cost of the other inputs used in constant proportion with the upstream supplier's input. If $d_1 < d_2$, then a discriminating upstream supplier will set prices such that $w_1 > w_2$ and $w_1 + d_1 < w_2 + d_2$.*

In words, this result says that the upstream supplier handicaps the lower-cost downstream firm, but by less than its initial efficiency advantage. A more general statement of Proposition 4 is proved in the Appendix.

This model is worth noting because it implies that large firms (in a Cournot model low-cost firms have high market shares) should oppose price discrimination, while small, inefficient firms should favor it. The observed pattern of support for antidiscrimination laws appears to be the opposite. To the extent that these laws deal with true price discrimination, this fact supports the view that the threat of backward integration, not differences in the efficiency of downstream firms, is the motivation for discrimination in intermediate goods markets.

There is a second type of welfare effect that may arise in a more general model. I have considered a fixed-proportions production technology for the final good. When factors may be used in variable proportions, changes in the price of an intermediate good lead to changes in the mix of inputs used by the downstream firms as they try to substitute away from those inputs whose relative prices have risen. In markets where the integration frontier is upward sloping, it is trivial to generalize the earlier welfare results under the usual assumptions that other factors are competitively supplied at constant marginal cost. The derived demands for the monopolistically supplied input still will be downward sloping and raising w_1 will lower the chain's profits, while raising w_2 will increase its profits. As before, price discrimination may lead to the upstream incumbent's increasing both prices. Each type of downstream firm will be driven to using an even more socially inefficient input mix. Thus, in these cases, there is an additional negative effect of price discrimination. In cases where price discrimination raises one price and lowers the other, one cannot in general determine whether the input distortions are made better or worse.

V. Conclusion

This analysis demonstrates that intermediate good and final good markets are fundamentally dissimilar, and that the conclusions derived from models of final good markets may be inappropriate for intermediate good markets. The fact that discrimination may raise prices charged to both types of buyers shows that the demand interdependencies and the possibility of integration that arise in intermediate good markets can have powerful effects on the equilibrium outcome.

Turning to policy implications, this analysis demonstrates that intermediate good price discrimination may shift prices in a way that reduces output in the final good market and thus lowers consumers' surplus and welfare. In other cases, price discrimination may increase welfare by preventing socially inefficient integration. These results suggest that there may be a useful role for government regulation of discriminatory pricing, but that a flat ban could have adverse welfare consequences. Unfortunately, the analysis does not reveal whether there is any implementable form of regulation that would be welfare improving.[13]

APPENDIX

Here, I generalize Lemma 1 and Propositions 1 and 4. The results proved depend only on those assumptions listed explicitly in the statements of the theorems.

There is an upstream monopolist who sells to n downstream firms in a representative market. If no downstream firm integrates, then the upstream supplier's profits are given by $U^m[\mathbf{w}]$, where $\mathbf{w}$ is the n-vector of input prices set by the supplier. Let Ω denote the set of all price vectors that do not induce integration by any downstream firm.

ASSUMPTION 1: *The downstream firms are Cournot competitors producing a homogeneous good with demand such that the industry marginal revenue curve is downward sloping.*

The restriction on demand guarantees that the downstream firms' second-order conditions are satisfied and that the Cournot equilibrium is unique.

ASSUMPTION 2: *Firm i has constant marginal costs of production equal to $d_i + w_i$ (d_i a nonnegative constant).*

ASSUMPTION 3: *If $(w, w,\dots, w) \in \Omega$, then for all λ such that $0 \le \lambda \le 1$ $(\lambda w, \lambda w,\dots, \lambda w) \in \Omega$.*

Assumption 3 is a weak restriction on the set of prices that prevent integration.[14] In particular, it allows for the threat of integration by several stores to be present (for example, the local store, as well as the chain, may have a credible threat of integration).

Let a caret over the variable denote the value that a variable takes when the upstream supplier is allowed to practice price discrimination. Some downstream firms may shut down under discrimination. For a firm that shuts down, define $\hat{w}_i$ to be the minimal price such that the firm remains inactive. Let A (for "active") denote the set of firms who have positive output levels under price discrimination. Let n_a denote the number of active firms. Finally, let w_a denote the average price charged by the upstream supplier to downstream firms who are active under discrimination:

$$\text{(A1)} \qquad w_a = \sum_{i \in A} \hat{w}_i / n_A.$$

LEMMA A1: *Under Assumptions 1–3, if $d_i = d$ for all i (d a nonnegative constant) and there is no integration under either pricing regime, then $w_a \ge w_b$, where w_b is the average price charged absent price discrimination.*

PROOF:

If $\hat{w}_i = \hat{w}_j$ for all i and j, then the upstream incumbent's ability to discriminate

[13] The present analysis does not provide support for the enforcement of the Robinson-Patman Act. F. M. Scherer (1980) provides a good survey of the analysis of the anticompetitive effects of the Act.

[14] One can drop all assumptions about the form of Ω by replacing Assumption 3 with the assumption that $U^m[w, w,\dots, w]$ is quasi concave when viewed as a function of the scalar w. The function $U^m[\cdot,\cdot,\dots,\cdot]$ will have this property when the conditions of Assumptions 1 and 2 are satisfied and demand for the final good is either linear or iso-elastic, for example. If one assumes directly that $U^m[\cdot,\cdot,\dots,\cdot]$ is quasi concave, then, as I show in my earlier paper, the Cournot assumption can be relaxed.

has no effect. Suppose that at least two prices differ from one another and consider the effects of the upstream supplier's charging w_a to firm i if $i \in A$ and $\hat{w}_i$ otherwise. As shown by Cournot, total output would remain at $\hat{X}$. Using the fact that $\hat{x}_i < \hat{x}_j$ if and only if $\hat{w}_i < \hat{w}_j$, it follows that revenues under this new set of prices, $w_a\hat{X}$, would exceed the original revenues, $\Sigma_{i=1}^{n}\hat{w}_i\hat{x}_i$. Lowering w_i to w_a for all firms would raise total output and (since $w_a \geq c$) yield even greater profits. Thus, $U^m[w_a, w_a, \dots, w_a] > U^m[\hat{w}_1, \hat{w}_2, \dots, \hat{w}_n] \geq U^m[w_b, w_b, \dots, w_b]$. Since w_a was not chosen as the nondiscriminatory price, it must not be in Ω. Hence, by Assumption 3, $w_a > w_b$.

One can now follow the line of argument made in the text, using the property that the Cournot equilibrium output level can be expressed as an increasing function of n_a and a decreasing function of w_a, to prove

PROPOSITION A1: *Under Assumptions* 1–3, *if $d_i = d$ for all i and there is no integration under either regime, then total output and total surplus are (weakly) lower when price discrimination is practiced than when it is forbidden.*

In the results above, the average price strictly rises and welfare strictly falls except in markets in which the upstream incumbent would not set discriminatory prices even if it were given the legal authority to do so.

PROPOSITION A4: *Under Assumptions* 1 *and* 2, *if Ω includes all possible price vectors, then a discriminating upstream supplier sets prices such that, if $d_i < d_j$, then $\hat{w}_i > \hat{w}_j$ and $\hat{w}_i + d_i < \hat{w}_j + d_j$.*

PROOF:

Suppose to the contrary that $\hat{w}_i \leq \hat{w}_j$ for some i and j such that $d_i < d_j$. Consider the effects of raising w_i by δ, where $0 < \delta < (\hat{w}_i + d_i - \hat{w}_j - d_j)/2$, while lowering w_j by δ. Let $\gamma > 0$ denote the amount by which x_i falls. x_j then rises by γ, and the other output levels are unaffected. Hence, the change in upstream profits is

(A2)
$$\delta\{(\hat{x}_i - \gamma) - (\hat{x}_j + \gamma)\} + \gamma\{\hat{w}_j - \hat{w}_i\} > 0,$$

which contradicts the optimality of $\hat{\mathbf{w}}$.

A similar argument shows that $\hat{w}_i - \hat{w}_j \geq d_j - d_i$ for some i and j where $d_i < d_j$ also leads to a contradiction. Therefore, the two conditions in the statement of the proposition must hold.

PROOF OF PROPOSITION 2:

(*ii*) $I[\cdot]$ *upward sloping*. First, suppose that changes in w_1 and w_2 have no effect on expected integration profits. For Cournot firms the lowest value of w_2 for which $x_2 = 0$ given w_1 solves $P[x(w_1, w_2)] = w_2$. Moreover, any higher level of w_2 will lead to the same equilibrium. When $x_2 = 0$, the slope of the supplier's iso-profit line is

(A3)
$$-\partial U^m/\partial w_1 \big/ \partial U^m/\partial w_2 = [-\{-x_1 + (w_2 - w_1)\partial x_1/\partial w_1 + (c - w_2)\partial X/\partial w_1\}] / [(w_2 - w_1)\partial x_1/\partial w_2 + (c - w_2)\partial X/\partial w_2].$$

(Here and below I am taking left-hand derivatives.) The slope of the chain's iso-profit line is given by

(A4)
$$-\partial \pi/\partial w_1 \big/ \partial \pi/\partial w_2 = [-\{-x_1 + (P - w_1)\partial x_1/\partial w_1 + x_1 P' \partial X/\partial w_1\}] / [(P - w_1)\partial x_1/\partial w_2 + x_1 P' \partial X/\partial w_2].$$

Using the fact that $w_2 = P$, both of these slopes can be expressed in the form

(A5)
$$H(\tau) \equiv \frac{-\{-x_1 + (w_2 - w_1)\partial x_1/\partial w_1 + \tau \partial X/\partial w_1\}}{(w_2 - w_1)\partial x_1/\partial w_2 + \tau \partial X/\partial w_2}$$

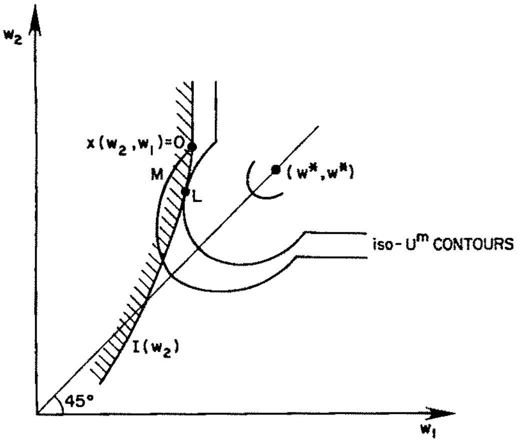


FIGURE 3

The chain chooses x_1 to satisfy the first-order condition $(P - w_1) + x_1 P' = 0$. Profit maximization by the upstream supplier implies $w_1 > c$. Hence, for $w_2 = P$, $x_1 P' = (w_1 - w_2) > (c - w_2)$. This argument demonstrates that the value of τ in equation (A4) is greater than its value in equation (A3).

Differentiation of equation (A5) shows that $H(\tau)$ is an increasing function of τ. Therefore, the slope of the chain's iso-profit line is greater than that of the upstream supplier's iso-profit line. By continuity, this relationship will hold in a neighborhood of the shutdown point, and it is not optimal for the supplier to induce firm 2 to shut down. Figure 3 illustrates this result. Point L yields higher upstream profits than does point M.[15] If π^e rises in response to an increase in either w_1 or w_2, then the slope of the chain's iso-profit line is even steeper, and the result continues to hold.

[15] The only relevant case is the one illustrated in Figure 3, where the upstream incumbent's iso-profit line has a positive slope. $\partial U^m/\partial w_2 < 0$ at point M, by the fact that $x_2 = 0$ and $w_1 < w_2$. If $\partial U^m/\partial w_1 < 0$, then the upstream supplier would reduce w_1. The possibility of being at the local store's shutdown point arises only when $\partial U^m/\partial w_1 > 0$, and equation (A3) is positive.

REFERENCES

Adelman, Morris A., "Dirlam and Kahn on the A&P Case," *Journal of Political Economy*, October 1953, *61*, 436–41.

Cournot, Augustin, *Mathematical Principles of the Theory of Wealth*, New York: Kelly, 1897.

Katz, Michael L., "The Welfare Effects of Third-Degree Price Discrimination in Intermediate Goods Markets," Woodrow Wilson School Discussion Paper in Economics No. 90, March 1985.

Milgrom, Paul and Roberts, John, "Limit Pricing and Entry Under Incomplete Information—An Equilibrium Analysis," *Econometrica*, March 1982, *50*, 443–59.

Ordover, Janusz A. and Panzar, John C., "On the Nonexistence of Pareto Superior Outlay Schedules," *Bell Journal of Economics*, Spring 1980, *11*, 351–54.

______ **and** ______, "On the Nonlinear Pricing of Inputs," *International Economic Review*, October 1982, *23*, 659–75.

Panzar, John C., and Sibley, David S., "Optimal Two Part Tariffs for Inputs: The Case of Imperfect Competition," unpublished draft, 1983.

Posner, Richard A., *The Robinson-Patman Act*, Washington: American Enterprise Institute for Public Policy Research, 1976.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

Schmalensee, Richard, (1981a) "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, *71*, 242–47.

______, (1981b)"Monopolistic Two-Part Pricing Arrangements," *Bell Journal of Economics*, Autumn 1981, *12*, 445–66.

Varian, Hal R., "Price Discrimination and Social Welfare," *American Economic Review*, September 1985, *75*, 870–75.

U.S. Department of Justice, *Report on the Robinson-Patman Act*, Washington: USGPO, 1977.

Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets

By PAUL L. JOSKOW*

This paper examines the importance of specific relationship investments in determining the duration of coal contracts negotiated between coal suppliers and electric utilities. Data for 277 coal contracts are used to perform the analysis. The results provide strong support for the view that buyers and sellers make longer commitments to the terms of future trade at the contract execution stage, and rely less on repeated bargaining, when relationship-specific investments are more important.

This paper seeks to test empirically the importance of relationship-specific investments in determining the duration of coal contracts negotiated between coal suppliers and electric utilities.[1] The analysis makes use of information for a large sample of coal contracts that were in force in 1979. It takes as a starting point Oliver Williamson's 1983 definitions and categorization of relationship-specific investments and applies them to the characteristics of coal market transactions. It also follows Williamson and Benjamin Klein, Robert Crawford, and Armen Alchian (1978) and assumes that risk aversion is not an important factor determining the structure of vertical relationships between coal suppliers and electric utilities.[2]

Coal market transactions are interesting to focus on because there is considerable variation in the duration and structure of vertical relationships between buyers and sellers. I observe spot market transactions, vertical integration, and a wide variety of longer-term contractual relationships with durations ranging from one year to fifty years.[3] My related work (1985) suggests that asset-specificity considerations may be an important factor affecting the structure of vertical relationships in coal markets. The empirical results reported below provide strong support for the hypothesis that buyers and sellers make longer *ex ante* commitments to the terms of future trade, and rely less on repeated negotiations over time, when relationship-specific investments are more important.

I. Contract Duration and Transaction-Specific Investments

The reliance on relationship-specific investments to support cost-minimizing ex-

*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. The preliminary version of this paper was written when I was a Fellow at the Center for Advanced Study in the Behavioral Sciences. Leslie Sundt and Rafe Leeman provided valuable research assistance. Keith Crocker, Henry Farber, Oliver Hart, Scott Masten, Jean Tirole, and Oliver Williamson read preliminary versions and made useful suggestions. Three anonymous referees provided helpful suggestions. I benefited from seminar presentations at UCLA, Stanford, and MIT. Support from MIT and the Center for Advanced Study in the Behavioral Sciences is gratefully acknowledged. The views expressed here and any remaining errors are my sole responsibility.

[1] Electric utilities account for over 80 percent of domestic coal consumption.

[2] Other empirical work in this tradition includes Kirk Monteverde and David Teece (1982) and Scott Masten (1984), which focus on vertical integration; Keith Crocker and Scott Masten (1986), J. Harold Mulhern (1986), and Victor Goldberg and John Erickson (1982) which focus on long-term contracts; my paper (1985) which examines both vertical integration and long-term contracts.

[3] About 15 percent of electric utility coal consumption is accounted for by transactions with integrated suppliers, 15 percent is accounted for by spot market purchases, and about 70 percent is accounted for by contracts with durations of one to fifty years. See my paper (1985, pp. 50–54).

change is frequently advanced as an important factor explaining why we observe the use of long-term contracts that establish the terms and conditions of repeated transactions between two parties *ex ante*.[4] According to transactions cost theory,[5] when exchange involves significant investments in relationship-specific capital, an exchange relationship that relies on repeated bargaining is unattractive. Once the investments are sunk in anticipation of performance, "hold-up" or "opportunism" incentives are created *ex post* which, if mechanisms cannot be designed to mitigate the parties' ability to act on these incentives, could make a socially cost-minimizing transaction privately unattractive at the contract execution stage.[6,7] A long-term contract that specifies the terms and conditions for some set of future transactions *ex ante*, provides a vehicle for guarding against *ex post* performance problems.[8]

A coal contract generally specifies in advance a method for determining the price that the buyer is obligated to pay for each delivery (generally a formula for determining prices for deliveries at each point in time),[9] quantities that the seller is obligated to deliver and the buyer is obligated to purchase at each point in time (usually monthly),[10] the quality of the coal (Btu, sulfur, ash, and chemical composition), the source of the coal, and the period of time over which the contractual provisions are to govern the terms and conditions of trade.[11] The primary readily quantifiable characteristics of coal contracts that appear to vary widely from contract to contract are the quantity and characteristics of the coal contracted for and the length of time that the parties agree *ex ante* to commit themselves to the terms and conditions specified in the contract. It is this length of time to which the parties agree *ex ante* to abide by the terms of a contract that I refer to as the "duration" of the contract.[12]

My hypothesis is that the more important are relationship-specific investments, the longer will be the period of time (or number of discrete transactions) over which the parties will establish the terms of trade *ex ante* by contract. I therefore expect to observe that the variation in the agreed upon duration of contractual commitments is directly related to variations in the importance of relationship-specific investments.[13]

[4] Williamson (1979, 1983), Klein et al., Oliver Hart and Bengt Holmstrom (1986, pp. 1–2; 86–101). Other reasons for the use of long-term contracts have also been suggested. These include information lags, income effects and risk aversion, and improved monitoring of performance.

[5] I use the term "transactions cost theory" to refer generally to the work of Williamson (1979, 1983, 1985) and Klein et al.

[6] As Klein et al. discuss, the sunk investments create a stream of quasi rents that gives one party or the other (or both) some *ex post* bargaining power.

[7] The presence of these contracting hazards and imperfections in the ability of the transacting parties to protect against them does not mean that a deal will not be made. It simply means that the costs of making the transactions—the cost of the coal in this paper—will be higher than it would be if these hazards could be fully mitigated. Both parties have an interest in trying to structure the relationship so that a cost-minimizing deal can be struck.

[8] As discussed in my earlier paper (1985), reputational considerations may provide a natural market constraint on "bad behavior" *ex post*. Reputational constraints reduce the need to write inflexible long-term contracts to support cost-minimizing exchange in the presence of asset specificity. Reputational constraints are likely to be imperfect in coal markets, however. At the other extreme, vertical integration may be chosen to deal with *ex post* performance problems if satisfactory contractual solutions cannot be found.

[9] See my earlier paper (1986).

[10] The typical contract specifies a monthly and annual delivery schedule subject to minimum and maximum production and take obligations. The allowed variations from the contracted quantities in actual contracts that I have reviewed is fairly small.

[11] There are many other provisions as well, including arbitration provisions, force majeur provisions, resale provisions, etc. These provisions are fairly standard in long-term coal contracts, however.

[12] The actual duration of a contract could be longer or shorter than this. Buyers and sellers frequently voluntarily negotiate an extension of an existing contract. Contracts may also be broken through breach or mutual agreement. As far as I can tell from the data that I have reviewed, however, coal contracts are rarely terminated prematurely See my paper (1986, p. 2).

[13] To the extent that there are tradeoffs between contract duration and the incidence and structure of other contractual "protective" provisions, such as the method for determining price adjustment and quantities, these other provisions should be included in the analysis as well. As indicated above, however, there

II. Asset Specificity and the Contractual Duration of Coal Supply Relationships

Williamson (1983, p. 526) identifies four distinct types of transaction-specific investments, three of which appear to be relevant to different types of coal supply relationships. The three types of relevance to coal market transactions are:[14]

(a) Site Specificity: The buyer and seller are in a "cheek-by-jowl" relationship with one another, reflecting *ex ante* decisions to minimize inventory and transportation expenses. Once sited the assets in question are highly immobile.

(b) Physical Asset Specificity: When one or both parties to the transaction make investments in equipment and machinery that involves design characteristics specific to the transaction and which have lower values in alternative uses.

(c) Dedicated Assets: General investment by a supplier that would not otherwise be made but for the prospect of selling a significant amount of product to a particular customer. If the contract is terminated prematurely, it would leave the supplier with significant excess capacity. Although Williamson does not discuss it, I think that there is probably a "buyer" side analogy to the dedicated asset story as well. A buyer that relies on a single supplier for a large volume of an input *may* find it difficult and costly to quickly replace these supplies if they are terminated suddenly and effectively withdrawn from the market and, as a result, a large unanticipated demand is suddenly thrown on the market.

As discussed in more detail in the Appendix, I have put together a data base that includes information for nearly 300 contracts between electric utilities and coal suppliers that were in force in 1979. The data base includes information of various kinds regarding the characteristics of the individual coal contracts, the suppliers, the buyers, and the quality and quantity of the coal contracted for. The strategy was to use the information about the individual contracts in the data base and to attempt to measure, at least ordinally, differences in the importance of transaction-specific investments of one or more of the types identified by Williamson for each contract.

Williamson's notion of "site specificity" is the easiest to capture explicitly for coal supply relationships. For most electric generating plants, coal is purchased in one of three major coal-producing regions and then transported by rail, barge, and/or truck (often at least two of these transport modes are involved) to the power plant where it is burned. However, there are a relatively small number of plants that have been sited next to specific mines in anticipation of taking all or most of their requirements from that mine. These "mine-mouth" plants are generally developed simultaneously with the mines themselves. This appears to be a classic case of the cheek-by-jowl relationship that Williamson has in mind when he discusses site specificity.[15] The potential for *ex post* opportunism problems arising if the parties were to rely on repeated bargaining appears to be especially great in this case.[16] I there-

appears to be relatively little variation in these provisions in my data base, especially relative to the very large variation in contract duration. I therefore feel that it is safe to assume for purposes of this analysis that we have a sample of contracts that essentially holds these other provisions constant. In any event, we can measure the utilization of these other provisions only for a small fraction of the contracts in the data base and therefore cannot examine such tradeoffs directly. Note that the coal market is not subject to the kinds of price regulation discussed in Masten-Crocker and Crocker-Masten regarding natural gas contracts.

[14] The fourth is what Williamson calls "human asset specificity" (1983, p. 526). Jean Tirole has suggested to me that Williamson's four types of relationship-(or transaction) specific investment are simply different instances of the same phenomenon. I believe that this is correct. However, I find the distinctions to be quite useful for empirical applications.

[15] This is discussed in much more detail in my 1985 paper.

[16] Williamson (1983) states that common ownership is the predominant response to site specificity. My work with coal supply arrangements indicates that common ownership (vertical integration) is much more likely to emerge for mine-mouth plants than other types of plants, but that contracts are also used to govern exchange for about half of the mine-mouth plants constructed since 1960. We would probably see more vertical integration

fore expect that contracts for supplies for mine-mouth plants will be much longer than the average contract involving supplies to other types of plants, other things equal.[17]

Let us turn next to physical asset specificity. When coal-burning plants are built, they are designed to burn a specific type of coal (see my 1985 paper; Richard Schmalensee and myself, 1986; my paper with Schmalensee, 1985). By "type" of coal, I mean coal with a specific Btu, sulfur, moisture, and chemical content. The type of coal that a generating unit is designed to burn affects its construction cost and its design thermal efficiency. Deviations from expected coal quality can lead to a deterioration in performance or require costly retrofit investments. Thus when a plant is designed, the operator becomes "locked in" to a particular type of coal.[18]

The fact that a plant is locked in to a particular type of coal does not necessarily imply that the buyer is locked in to a specific supplier, however. Whether or not the plant design/coal characteristic lock in also leads to a lock in with the current supplier depends on other characteristics of the transaction. In particular, it is likely that the relationship between this type of asset-specificity and *ex post* hold up or opportunism problems is related to inter- and intraregional variations in coal quality, least cost supply technology, and transportation alternatives.[19]

The characteristics of coal produced in the United States vary systematically among the three major coal-producing regions. The eastern coal-producing districts produce high Btu coal of reasonably uniform quality. The midwestern coal-producing districts produce lower Btu coal that generally has a very high sulfur content. Coal quality is also more variable than that in the East. Finally, the western coal-producing districts generally produce coal with a much lower Btu content and a very low sulfur content. The quality of the coal varies quite widely throughout the western region.[20]

In addition to variations in coal quality among the regions, there are also systematic variations in the least-cost technology for producing coal and in the transportation alternative available. In the East, relatively small underground mines are economical and the supply of eastern coal can be expanded fairly quickly. Relatively abundant transportation alternatives, combined with relatively short average transport distances, mean that transportation is not likely to be a significant barrier to a buyer's obtaining alternative supplies. In the West, large surface mines that can be most economically expanded in large "lumps," are the least-cost production technology. Transportation alternatives are poor, the average transport distance quite long, large unit train shipments are the most economical transport method,[21] and utilities often must rely on one or two railroads to move the coal. The situation in the Midwest lies somewhere between these two extremes.[22]

for mine-mouth plants if state and federal regulation of electric utilities did not discourage it. I have argued elsewhere (1985) that to the extent that electric utility regulation biases coal supply arrangements at all, it is probably to make short-term purchases more desirable than they might otherwise be. While utilities might also like to integrate backwards into coal production to shift profits from a regulated to an unregulated activity, the regulatory process has discouraged this.

[17] I have included two plants in this category that are not technically mine-mouth plants but have economic characteristics that are identical to those of mine-mouth plants. For example, if a mine and a plant are connected by a transportation facility (a slurry pipeline or a rail line) built and owned by the supplier or buyer specifically to transport coal from a specific mine to a specific plant, the associated coal contract was grouped with the mine-mouth plants.

[18] Exactly how locked in, is a variable of choice, however.

[19] Transportation costs are on average a large fraction of delivered costs and lining up efficient transportation arrangements for large quantities of coal can be a time-consuming process.

[20] The midwestern region is sometimes broken up into two subregions (eastern and western interior) in discussions of coal supply. The western region is sometimes broken up into three or more subregions. Texas, where lignite coal is produced, is often considered a separate producing region. My data base has no contracts for Texas coal and I do not discuss that area here.

[21] Unit train cars are often owned or leased by the utility rather than by the railroad.

[22] See Martin Zimmerman (1981, pp. 17–36).

There are also systematic differences in the relative and absolute importance of spot markets in the three supply regions.[23] On average, from 1974 through 1982 spot market transactions accounted for roughly 15 percent of total domestic coal purchases by electric utilities. In 1982, spot market sales accounted for about 10 percent of coal supplies or about 60 million tons. However, in the western region, less than 2 percent of the coal delivered to electric utilities was purchased on the spot market or less than 5 million tons.[24] The spot market is more active in the Midwest, accounting for about 8 percent of deliveries in 1982 or about 10 million tons per year. The spot market is most active in the East where about 18 percent of deliveries went through the spot market in 1982 or about 45 million tons.[25]

These considerations imply the following: Coal suppliers are likely to be less able to exploit the lock-in effect associated with boilers designed to burn coal with specific characteristics in the East than in the West.[26] Thus the protection of a long-term contract is likely to be more desirable from the buyers' (and the sellers'—see below) perspective, for transactions involving western coal relative to transactions involving eastern coal, with midwestern coal falling somewhere in between.

Finally, let us turn to "dedicated asset" considerations. The available information in the data base does not make it possible for us to know specifically whether the supplier made general investments that would not have been made but for the prospect of selling a *significant amount of product* to a particular customer and if the contract is terminated prematurely it would leave the supplier with excess capacity (see Williamson, 1983, p. 526).[27] Williamson's conceptualization of dedicated assets implies that the importance of this factor in structuring coal supply relationships should vary with the quantity of coal that is initially contracted for, other things equal. The larger the annual quantity of coal that is contracted for, the more difficult it is likely to be for the seller to quickly dispose of unanticipated supplies (if the buyer breaches) at a compensatory price, and the more difficult will it be for a buyer to replace supplies at a comparable price if the seller withdraws them from the market. Thus, I expect that the greater the annual quantity of coal contracted for the longer will be the specified duration of the contract.

Because of systematic variations in the optimal scale and capital intensity of coal production across regions, dedicated asset considerations are also likely to be more important for western coal than for eastern coal. The greater heterogeneity in coal supplies and the difficulties of obtaining suitable transportation for it in the West, suggest that dedicated asset problems are likely to be more severe in the western than the eastern region. The very thin spot market in the

[23] Spot market sales also vary from year to year. Spot market transactions tend to be higher when coal miner strikes are anticipated as utilities seek to build up stockpiles or after coal mining strikes are over and stockpiles are replenished. The volume of spot market transactions also varies in response to unanticipated changes in coal supply and demand.

[24] The aggregate volume of spot market transactions for western coal is quite small compared to the annual quantity of western coal contracted for in a typical contract. The contracts for western coal in my data base have a mean quantity of about 1.8 million tons per and a maximum quantity of over 8 million tons per year.

[25] I believe that the wide variation in the importance of the spot market in different regions is largely related to the same economic considerations that lead me to conclude that the importance of asset specificity also varies from region to region.

[26] Generating plants located along the eastern seaboard that use coal essentially always use eastern coal. These plants are also more likely to have units that have multifuel capabilities than plants located elsewhere and can switch back and forth between coal and oil or gas (often with some performance penalty). See my paper with Frederick Mishkin (1977). Purchasers of eastern coal with multifuel capabilities will be less susceptible to opportunism problems when coal and oil prices are close together.

[27] Indeed, as a practical matter, it is unclear to me how one could ever go about determining this directly given the data that are likely to be available for analysis. The coal contracts that I have reviewed do sometimes have language that appears to recognize the dedicated asset notion directly, but the absence of an explicit statement cannot be assumed to imply that dedicated asset considerations are not important.

western region should be especially problematical for both sellers and buyers when large contractual commitments are breached because of the heterogeneity of the coal, the characteristics of least-cost production and the limited transportation alternatives. This all implies again that contracts for western coal should have longer contractual durations than contracts for eastern coal.

To summarize, if variations in the importance of relationship-specific investments do in fact lead to variations in the extent to which the parties precommit to the terms of future trade *ex ante*, I expect to find that the duration of contractual relationships specified at the contract execution stage will vary systematically with three primary observable characteristics of coal supply transactions. First, whether the plant taking the coal is a mine-mouth plant or not. I expect to observe longer-term contracts negotiated for mine-mouth plants. Second, with the region of the country in which the coal is produced. I expect the western region to have the longest contracts and the eastern region the shortest with the midwestern region having contracts with prespecified durations that lie in between. Third, with the annual quantity of coal contracted for. I expect that coal supply arrangements involving large annual quantity commitments will be supported by longer contracts than supply arrangements involving smaller quantities of coal.

III. Model Specification and Estimation

I am primarily interested in estimating a set of simple relationships between the duration of contractual commitments (*DURATION*) specified by the parties at the contract execution stage and, (a) the *annual* quantity (*QUANTITY*)[28] of coal contracted for, (b) a dummy variable (*MINE-MOUTH*) that takes on a value of 1 for a mine-mouth plant and zero otherwise, and (c) dummy variables that indicate the coal supply region in which the supplier is located (*MIDWEST* and *WEST*, so that regional effects are measured relative to contracts for eastern coal). Additional variables are considered in the next section.

The data base that I use includes information for approximately 300 coal supply contracts between domestic coal suppliers and investor-owned electric utilities. The contracts included in the data base were negotiated in various years up through 1979 and were in force at least for part of 1979. The data are discussed in more detail in the Appendix. Information on all variables of primary interest for this study is available for 277 of the contracts in the data base. I present estimates using both the full 277 observation sample as well as a subsample consisting of 169 contracts that involve deliveries dedicated to a single power plant.[29] Table 1 provides the mean, standard deviation, minimum and maximum values, and a brief description for all of the variables used in this section and subsequent sections for both the primary sample and the single-plant subsample. Table 2 is a correlation matrix for all of the variables in the two samples with the correlations for the 277 observation sample below the diagonal and those for the 169 observation subsample above the diagonal.

I work first with three simple specifications of the contract duration equation:

$$(1)\quad DURATION_i = a_0 + b_1 QUANTITY_i + b_2 QUANTITY_i^2 + b_3 MINE\text{-}MOUTH_i + b_4 MIDWEST_i + b_5 WEST_i + u_i$$

[28] Quantities are expressed in terms of the thermal (Btu) content of the coal. The basic results are not affected when quantities are expressed in tons. For the extensions reported in the next section, normalizing quantities for Btu content makes it possible to more accurately examine the effects, if any, of contract quantities relative to total plant and utility utilization of coal.

[29] Several people suggested to me that relationship-specific investment effects are most likely to be revealed for contracts that dedicate all supplies to a single plant. I also focus on this subsample to obtain the data necessary to explore issues discussed in the next section.

TABLE 1—SAMPLE STATISTICS

Variable	Observations	Description	Mean	Minimum	Maximum	Standard Deviation
DURATION	277	Contract Duration	12.75	1.00	50	10.43
	169	(years)	14.18	1.00	43	10.77
QUANTITY	277	Annual Contract	20.45	0.3696	183.00	24.62
	169	Quantity (trillion Btu's)	22.83	0.3696	183.00	27.05
PLANT PROPORTION	169	Fraction of Total Plant Use from Contract	0.44	0.03	1.00	0.35
UTILITY PROPORTION	169	Fraction of Total Utility Coal Use from Contract	0.19	0.003	1.00	0.23
PLANT QUANTITY	169	Plant Utilization of Coal (trillion Btu's)	51.47	2.95	172.42	41.69
UTILITY QUANTITY	169	Utility Utilization of Coal (trillion Btu's)	221.54	2.95	919.80	270.56
PLANT/UTILITY	169	Plant Use as Fraction of Total Utility Use	0.455	0.007	1.00	0.323
MINE-MOUTH	277	Mine-Mouth Plant		($D=1$; 14 Observations)		
	169	Dummy Variable		($D=1$; 14 Observations)		
WEST	277	Western Region		($D=1$; 54 Observations)		
	169	Supply Dummy		($D=1$; 44 Observations)		
MIDWEST	277	Midwestern Region		($D=1$; 68 Observations)		
	169	Supply Dummy		($D=1$; 47 Observations)		
DATE-71	277	Contracts Signed		($D=1$; 43 Observations)		
	169	1971–73: Dummy		($D=1$; 29 Observations)		
DATE-74	277	Contracts Signed		($D=1$; 116 Observations)		
	169	1974–77: Dummy		($D=1$; 71 Observations		
DATE-78	277	Contracts Signed		($D=1$; 72 Observations)		
	169	1978–79: Dummy		($D=1$; 38 Observations)		
YEAR	277	Year Contract Executed	1974	1955	1979	4.49
	169		1974	1955	1979	4.64

[a] Data sources and variable definitions can be found in the Appendix.
[b] The 169 observation subsample includes contracts dedicated to a single plant.

TABLE 2—CORRELATION MATRIX

277 Observation Sample	169 Observation Sample											
	DURATION	*QUANTITY*	*LOG-QUANTITY*	*MINE-MOUTH*	*MIDWEST*	*WEST*	*YEAR*	*PLANT PROP.*	*UTILITY PROP.*	*PLANT/UTILITY*	*PLANT QUANTITY*	*UTILITY QUANTITY*
DURATION	–	0.60	0.68	0.64	0.004	0.51	−0.57	0.58	0.43	0.05	0.29	0.02
QUANTITY	0.60	–	0.78	0.42	−0.04	0.28	−0.35	0.48	0.43	0.05	0.41	0.11
LOG-QUANTITY	0.64	0.80	–	0.30	0.02	0.32	−0.44	0.57	0.42	0.02	0.45	0.17
MINE-MOUTH	0.54	0.38	0.27	–	−0.09	0.41	−0.37	0.46	0.41	0.05	0.10	−0.05
MIDWEST	0.14	0.08	0.11	−0.06	–	−0.37	−0.20	0.04	−0.04	−0.06	−0.03	−0.15
WEST	0.41	0.25	0.26	0.39	−0.28	–	−0.10	0.46	0.40	0.16	−0.06	0.16
YEAR	−0.63	−0.39	−0.44	−0.31	−0.27	−0.08	–	−0.28	−0.11	0.13	−0.16	−0.11
PLANT PROP.	–	–	–	–	–	–	–	–	0.61	−0.11	−0.24	−0.14
UTILITY PROP.	–	–	–	–	–	–	–	–	–	0.66	−0.08	−0.36
PLANT/UTILITY	–	–	–	–	–	–	–	–	–	–	0.15	−0.57
PLANT QUANTITY	–	–	–	–	–	–	–	–	–	–	–	0.34
UTILITY QUANTITY	–	–	–	–								–

Note: Figures below the diagonal are for the 277 observation sample, those above the diagonal are for the 169 observation subsample.

$$(2)\quad DURATION_i = a_0 + b_1 LOG\text{-}QUANTITY_i + b_3 MINE\text{-}MOUTH_i + b_4 MIDWEST_i + b_5 WEST_i + u_i$$

$$(3)\quad \log(DURATION_i) = a_0 + b_1 LOG\text{-}QUANTITY_i + b_3 MINE\text{-}MOUTH_i + b_4 MIDWEST_i + b_5 WEST_i + \log(u_i)$$

where i indexes contracts and u_i is an error term whose characteristics will be discussed further below.

I have allowed (*QUANTITY*) to enter these relationships nonlinearly by introducing a quadratic in quantity (*QUANTITY-SQUARED*) in (1) and using the natural logarithm of quantity (*LOG-QUANTITY*) in equations (2) and (3). Since powerplants have useful lives of roughly forty years and the costs of breach are likely to decline over time as plants and mines age, I expect that the impact of quantity on contractual duration will diminish as quantity increases. The following pattern of coefficient estimates for the three equations is implied by the hypothesized relationship between asset specificity and contract duration:

(*i*) All of the b_i's should be positive, except for b_2, which could be positive, negative, or zero (no nonlinearity), although I expect that it will be negative.

(*ii*) b_4 should be smaller than b_5.

Equations (1), (2), and (3) are estimated in three different ways. First, I present ordinary least squares (OLS) estimates of each equation for both samples. Next, I present OLS estimates that introduce dummy variables which indicate the date that the contracts were executed. Finally, I present maximum likelihood estimates based on the assumption that we have a truncated sample drawn from a population with either a normal or a log-normal density function. I discuss the rationale and results for each estimation approach and also present OLS results for contracts with suppliers in each of the three regions.

A. *OLS Estimates*

If we assume that the u_i in (1), (2), and (3) are independently distributed and drawn from a normal distribution with mean zero, then OLS will yield an unbiased estimator of the coefficients of interest. I proceed first with this assumption and provide estimates for alternative assumptions about the error structure below. The OLS results are presented in Table 3. The first three columns are estimates for the three equations using the 277 observation sample. Columns 4, 5, and 6 contain estimates for the 169 observation subsample.

The OLS estimates are, in all cases, consistent with the hypothesized relationship between asset specificity and contract duration. The effects of annual contract quantity, region, and mine-mouth plants have the predicted signs and are estimated quite precisely. A mine-mouth plant is predicted to have a contract that is about 16 years longer than those of other plants. Contracts with eastern producers are 3 to 5 years shorter than those for western and midwestern producers. Contracts with western producers are 2 to 3 years longer than those with midwestern producers. The difference in duration between the *WEST* and the *MIDWEST* is generally not significant at the 5 percent level for the 277 observation sample, but is significant for the single-plant subsample. An increase in annual contract quantity of 22 trillion Btu's (roughly 1 million tons) yields about a 13-year increase in contract duration. As a general matter, the estimates are more precise for the sample of contracts dedicated to a single plant.

B. *OLS Estimates with Contract Date Dummies*

The desirable properties of the OLS estimates depend on the strong assumptions made about the error structure. Since the contracts in the data base were signed at many different times, it is natural to consider the possibility that contracting practices

TABLE 3—CONTRACT DURATION[a]

	277 Sample			169 Sample					*2SLS*
			LOG-			*LOG-*			Estimate
Independent	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	Duration
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
QUANTITY	0.4289	–	–	0.4091	–	–	–	–	–
	(0.0373)			(0.0040)					
QUANTITY-SQUARED	−0.0024	–	–	−0.0020	–	–	–	–	–
	(0.00030)			(0.00003)					
LOG-QUANTITY	–	4.4206	0.5057	–	4.2080	0.4942	4.2022	4.2057	5.1066
		(0.3742)	(0.0425)		(0.4069)	(0.0453)	(0.4617)	(0.4084)	(0.812)
MINE-MOUTH	16.3300	16.4317	0.5104	15.9583	16.2300	0.4616	16.3432	16.2284	15.4391
	(2.0496)	(2.0045)	(0.2279)	(1.9106)	(1.8421)	(0.2050)	(1.9426)	(1.8477)	(1.968)
MIDWEST	3.4267	3.8795	0.5154	2.7832	2.7843	0.5785	2.7317	2.7848	2.4268
	(0.9682)	(0.9821)	(0.1116)	(1.0928)	(1.1032)	(0.1228)	(1.1295)	(1.1065)	(1.153)
WEST	5.3550	5.2033	0.6142	5.9856	5.6108	0.6844	5.6456	5.6391	4.8916
	(1.357)	(1.1641)	(0.1323)	(1.2346)	(1.2586)	(0.1401)	(1.3406)	(1.2751)	(1.394)
PLANT PROPORTION	–	–	–	–	–	–	0.9729	–	–
							(1.9806)		
UTILITY PROPORTION	–	–	–	–	–	–	−2.0570	–	–
							(2.5832)		
PLANT/ UTILITY	–	–	–	–	–	–	–	−0.2246	–
								(1.4265)	
Constant	3.6770	−0.7902	0.6014	3.9334	0.0155	0.6242	−0.0146	0.1157	−1.8922
	(0.6586)	(0.9579)	(0.1089)	(0.8109)	(1.0917)	(0.1215)	(1.0978)	(1.2665)	(1.852)
Corrected R-squared	0.61	0.60	0.51	0.71	0.70	0.61	0.70	0.70	–
Observations	277	277	277	169	169	169	169	169	169

[a] OLS estimates. Standard errors of coefficient estimates are shown in parentheses.

changed over time. Not only might the duration of a typical contract have changed over time, but such changes may have been correlated with changes in contract quantities, supply location, and the development of mine-mouth plants over time. Failing to include variables indicated contract dates could then lead to a correlation between the independent variables and the error term. The OLS estimates would then be biased. To check to see if the estimates are sensitive to the presence of a left-out variable reflecting the contracting date, in Table 4, I report estimates of equations (1), (2), and (3) that have contract date dummies included. These contract date dummy variables are *DATE-71*, which is equal to one for all contracts signed between 1971 and 1973 inclusive, *DATE-74*, which equals one for contracts signed between 1974 and 1977, and *DATE-78* which equals one for contracts signed in 1978 and 1979. Since a separate variable for contracts signed prior to 1971 is not included, the coefficient estimates are all relative to pre-1971 contracts (i.e. the constant term). This aggregation of signing dates was made to reflect major shocks to coal supply and/or demand.[30]

[30] The 1971–73 period is just after the Clean Air Act Amendments of 1970 were passed, the 1974–77 period is the period after the Arab oil embargo and includes the subsequent increases in fossil fuel prices. The 1978–79 period coincides with the beginning of a slowdown in utility capacity additions. These periods are discussed in more detail in my paper (1986). The aggregation chosen is the same used to analyze pricing behavior in that paper. The results reported here are not sensitive to this aggregation, however. The same qualitative results are obtained if separate dummy variables are used for each year during the 1970's plus a separate dummy variable for pre-1965 and 1966–70 contracts.

TABLE 4—CONTRACT DURATION[a]

			LOG-			*LOG-*		
Independent	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
QUANTITY	0.3120	–	–	0.3355	–	–	–	–
	(0.03547)			(0.0406)				
QUANTITY-SQUARED	−0.0018	–	–	−0.0018	–	–	–	–
	(0.00027)			(0.00029)				
LOG-QUANTITY	–	3.0482	0.3245	–	3.3485	0.3631	3.2847	3.3461
		(0.3655)	(0.0380)		(0.4330)	(0.0446)	(0.4923)	(0.4344)
MINE-MOUTH	13.9437	13.6701	0.3140	14.6494	14.6907	0.3792	14.5742	14.6800
	(1.8482)	(1.8260)	(0.1899)	(1.8495)	(1.8347)	(0.1891)	(1.9423)	(1.8403)
MIDWEST	1.6814	1.9761	0.3029	1.4906	1.6405	0.4083	1.5543	1.6323
	(0.8785)	(0.8952)	(0.0931)	(1.0544)	(1.0786)	(0.1112)	(1.1083)	(1.0821)
WEST	4.8429	4.8662	0.4831	5.1054	5.1731	0.5137	5.0697	5.1258
	(1.0301)	(1.0549)	(0.1097)	(1.2183)	(1.2524)	(0.1291)	(1.3338)	(1.2676)
PLANT PROPORTION	–	–	–	–	–	–	0.9930	–
							(1.8907)	
UTILITY PROPORTION	–	–	–	–	–	–	−0.9138	–
							(2.4975)	
PLANT/UTILITY	–	–	–	–	–	–	–	0.3785
								(1.3660)
CONSTANT	12.0145	9.4184	1.7205	9.2341	6.1982	1.4220	6.1583	6.0806
	(1.2526)	(1.5307)	(0.1592)	(1.5185)	(1.8612)	(0.1918)	(1.8853)	(1.9142)
DATE-71	−2.3734	−2.4564	−0.0988	−0.7282	−0.9103	−0.0311	−0.9389	−0.9455
	(1.2715)	(1.2876)	(0.1339)	(1.4890)	(1.5057)	(0.1552)	(1.5290)	(1.5154)
DATE-74	−6.6815	−7.3044	−0.5098	−3.9714	−4.3786	−0.3149	−4.3848	−4.4134
	(1.2647)	(1.1446)	(0.1190)	(1.3483)	(1.3685)	(0.1410)	(1.3789)	(1.3781)
DATE-78	−10.5052	−10.6151	−1.3926	−7.0789	−6.5193	−1.0317	−6.4995	−6.5749
	(1.2647)	(1.2976)	(0.1349)	(1.5540)	(1.6324)	(0.1682)	(1.6702)	(1.6493)
Corrected *R*-squared	0.70	0.69	0.67	0.74	0.73	0.69	0.73	0.73
Observations	277	277	277	169	169	169	169	169

[a] OLS/Contract date dummies. Standard errors of coefficient estimates are shown in parentheses.

While the introduction of these contract date dummy variables may help to control for contract date related correlations between the error term and the independent variables, the estimated coefficients of these variables themselves have no obvious economic meaning. This is because of the nature of the sample. Recall that I observe contracts *in force* in 1979. If we think of the population as consisting of contracts written for particular plants (i) in a particular year (t), we can observe a contract only if

$$(4)\quad DURATION_{it} \geq (1979 - \text{Contract } YEAR)$$

This means that of those contracts signed in 1970, I can observe, in 1979, only those that had durations of at least 9 years, while I will observe shorter contracts that were signed in later years. Even if there were no changes in contracting behavior over time, we would inevitably find that the coefficients of the contract date dummies indicate that the average length of a contract in the sample is negatively correlated with the date of the contract.[31] The coefficients of the contract date variables can therefore tell us nothing directly about the changes in contracting behavior over time.

With these considerations in mind, we can turn to the results reported in Table 4, columns 1 through 6. The results obtained are again consistent with the hypothesized relationship between asset specificity and contract duration. The coefficients of the quantity, mine-mouth, and regional variables

[31] In the sample, the simple correlation between contract date (*YEAR*) and *DURATION* is about −0.60. See Table 2.

continue to be of the predicted signs and relative magnitudes. The only interesting difference between these results and those reported in Table 3 is that the difference between the durations of contracts signed with western and midwestern producers is now generally statistically significant at the 5 percent level for both samples.

C. *Maximum Likelihood Estimates*

A third alternative for estimating these equations is to follow Keith Crocker and Scott Masten (1986), who work with a sample of natural gas contracts with similar sampling properties, and assume that the sampling procedure which chooses contracts in force in a single year (1981 in their paper) represents a classical sample truncation problem as discussed by G. S. Maddala (1983, pp. 165–170). The population of contracts then implicitly consists of all contracts written since the earliest contract in the data base. We obtain a truncated sample because we observe contracts only if the duration of the contract is greater than or equal to L_i, where L_i is equal to 1979 minus the contract date. In this case, OLS estimates of (1), (2), and (3) would be biased, because the sampling process is likely to induce a correlation between the independent variables and the error term.

We can obtain estimates with desirable asymptotic properties by specifying the likelihood function of the sample, given the nature of the sampling truncation, and then solve for the maximum likelihood estimates (MLE) of the coefficients of interest. Following Maddala (p. 166), I assume that the population relationship between contract duration and the independent variables has a normally distributed error and that each observation is truncated at L_i. This leads to a standard maximum likelihood estimator.

The maximum likelihood estimates are reported in Table 5, columns 1 through 6.[32] The results are again quite consistent with the hypothesized relationship between asset specificity and contract duration. The signs and magnitudes of the coefficients of *QUANTITY*, *MINE-MOUTH*, *WEST*, and *MIDWEST* are again as predicted. With the exception of the mine-mouth dummy in equation (3), the coefficients are estimated quite precisely. The magnitudes of the estimated coefficients in this table are difficult to compare directly with those in Tables 3 and 4 because of the need to incorporate the truncation effects into estimates of contract duration.[33] Correcting for the sample truncation, the estimates reported in Table 5, column 5, for example, yield the following expected durations: The expected duration evaluated at the means of the independent variables is 10.5 years (compared to a mean of 14.2 years for the truncated 169 observation sample). Mine-mouth plants have contracts with an expected duration that is about 12 years longer than the average non-mine-mouth plant. Midwestern contracts are about 3.5 years longer than eastern contracts. Western contracts are about 11 years longer than eastern contracts and 6 years longer than midwestern contracts. The difference between western and midwestern contracts is generally significant at the 5 percent level.[34]

D. *Estimates for Individual Coal Supply Regions*

Finally, in Table 6, I report estimates of the relationship between *DURATION*, *LOG-QUANTITY*, and *MINE-MOUTH* for

[32] The estimates were obtained using the MLE routine in the *Statistical Software Tools* package (Version 1.0 as of October 1986) developed by Jeffrey Dubin and R. Douglas Rivers running on an IBM XT. The estimates in cols. 3 and 6 assume that the density function is log-normal. Contracts executed in 1979 have been dropped since the likelihood function includes terms that require taking the logarithm of $L_i = (1979 - YEAR)$ which is zero for contracts written in 1979. Since the shortest contract in the data base has a duration of one year, we can impose a lower bound between zero and 1 on L_i to include all observations. Estimates obtained using different lower bounds does not change the results, so I simply report results for the sample excluding the small number of contracts in the data base executed in 1979.

[33] See Maddala (p. 167).

[34] I have also produced maximum likelihood estimates of equation (2) for subsamples consisting of pre-1974 contracts and those signed between 1974 and 1979. The hypothesized relationship between contract duration and the variables representing variations in asset specificity persists for both subsamples.

TABLE 5—CONTRACT DURATION[a]

									2SLS
			LOG-			*LOG-*			*ESTIMATE*
Independent	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*	*DURATION*
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
QUANTITY	0.7948	–	–	0.5807	–	–	–	–	–
	(0.1046)			(0.0717)					
QUANTITY-SQUARED	−0.0043	–	–	−0.0030	–	–	–	–	–
	(0.00061)			(0.0004)					
LOG-QUANTITY	–	11.8527	0.6733	–	8.4367	0.6340	8.2709	8.4505	8.5737
		(1.5663)	(0.0897)		(1.0816)	(0.0816)	(1.0854)	(1.0815)	(3.5777)
MINE-MOUTH	16.5557	13.8763	0.2407	15.5484	14.4004	0.2736	14.1237	14.3841	15.4508
	(4.2407)	(3.8955)	(0.5714)	(2.9435)	(2.7764)	(0.4614)	(3.0120)	(2.7650)	(2.1083)
MIDWEST	8.3650	8.1377	0.4493	5.0050	4.7042	0.5376	4.4397	4.7468	2.4404
	(2.8188)	(2.5461)	(0.2281)	(2.5018)	(2.4511)	(0.2763)	(2.5221)	(2.4820)	(2.1797)
WEST	15.4864	13.8045	1.0500	11.7771	10.5506	1.0522	10.1238	10.4730	6.9477
	(4.1940)	(3.4244)	(0.0500)	(3.0124)	(2.4220)	(0.2416)	(2.6136)	(2.4095)	(1.8329)
PLANT PROPORTION	–	–	–	–	–	–	2.4401	–	–
							(3.7927)		
UTILITY PROPORTION	–	–	–	–	–	–	−2.0043	–	–
							(3.1937)		
PLANT/ UTILITY	–	–	–	–	–	–	–	1.4626	–
								(2.1945)	
Constant	−18.4532	−35.6427	−0.4699	−6.9510	−19.0118	−0.2789	−19.0377	−19.7755	−36.1521
	(5.7203)	(7.2690)	(0.26709)	(3.4094)	(4.6758)	(0.2448)	(4.8545)	(4.8645)	(6.9717)
Log-Likelihood	−781.08	−769.29	−174.29	−475.17	−466.99	−101.60	−466.68	−466.83	−451.38
Observations	277	277	255	169	169	160	169	169	169

[a] Maximum likelihood estimates. Standard errors of coefficient estimates are shown in parentheses.

TABLE 6—CONTRACT DURATION[a]

	Dependent Variable: *DURATION*							
Independent	*WEST*	*MIDWEST*	*EAST*	*EAST*	*WEST*	*MIDWEST*	*EAST*	*EAST*
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LOG-QUANTITY	3.4825	5.1249	4.2362	4.2968	3.5878	4.0269	4.408	4.4992
	(1.0714)	(0.7646)	(0.4668)	(0.4661)	(1.0992)	(0.6538)	(0.5798)	(0.5786)
MINE-MOUTH	18.4179	11.8107	–	17.6804	17.616	12.8372	–	17.0887
	(2.5064)	(5.1695)		(3.7348)	(2.3183)	(3.6363)		(3.8111)
Constant	6.819	1.365	−0.3975	−0.5265	7.2101	3.4075	0.4106	−0.6028
	(3.1610)	(5.1695)	(1.1159)	(1.1151)	(3.3441)	(1.8264)	(1.4256)	(1.4246)
Corrected R-squared	0.68	0.42	0.35	0.43	0.73	0.52	0.42	0.54
Observations	54	68	155	158	44	47	78	81

[a] OLS by region. Standard errors of coefficient estimates are shown in parentheses.

each coal supply region.[35] The samples do not include any mine-mouth plants using eastern coal, so columns 3 and 7 simply report the estimated relationship between *DURATION* and *LOG-QUANTITY*. There are in fact a few mine-mouth plants in the East and I have some information for three of them. These were not included in the sample because the data base did not have information on annual contract quantities for them. However, I was able to obtain information for delivered quantities for three eastern mine-mouth plants and have aug-

[35] I report only this variant of equation (2) to conserve space. It should be clear by now that the alternative specifications of (1), (2), and (3) do not yield any important differences in results.

mented the sample to include these plants, using delivered quantities rather than contract quantities as the values for the *QUANTITY* variable. These results are reported in columns 4 and 8 of Table 6.[36]

The effects of contract quantity and the mine-mouth dummy on contract duration are clearly not simply associated with the contracting behavior for coal from a particular region. The expected effects are found in each of the three regions. The coefficients of *QUANTITY* and *MINE-MOUTH* are of the expected signs and are estimated precisely in all cases. While there are differences in the magnitudes of the coefficients of these variables between the three regions, they are not very large numerically and equality of the coefficients of *QUANTITY* and *MINE-MOUTH* across regions cannot be rejected at standard significance levels. Contracts basically simply get longer as we move from East to West, other things equal.

IV. Alternative Measure of Asset Specificity

Clearly, the hypothesized relationship between contract duration and the variables that I have chosen to capture variations in asset specificity is quite robust to alternative specifications, samples, and estimating technique. Nevertheless, it is natural to ask whether there are alternative or additional factors that might explain the observed variations in, contract duration. One argument that has been suggested to me is that it is not only the size of the contractual commitment that is likely to be important, but also the fraction of a plant's, and perhaps the utility's, requirements obtained from a specific contract. The argument is that as a larger fraction of a plant's requirements is associated with a specific supplier, "physical asset specificity" attributes are likely to be more important and lead to longer contracts. It has also been suggested to me that opportunism problems are likely to be less severe if the utility as a whole is not heavily dependent on supplies provided pursuant to a specific contract. These arguments imply that variables measuring plant and/or utility "dependence" on a specific contract should be introduced into the contract duration relationship.

To examine this possibility, I have included variables in (2)[37] that measure the fraction of a plant's requirements (*PLANT PROPORTION*) and the fraction of the total coal requirements of the utility (*UTILITY PROPORTION*) that operates the plant which are accounted for by a particular contract. As an alternative, I also estimate (2) introducing a variable that is equal to total plant utilization of coal divided by total utility utilization of coal (*PLANT/UTILITY*). I can estimate these relationships only for the contracts that are for delivery to a single plant because it is only for these contracts that I can construct a meaningful measure of plant specific dependence (i.e., the 169 observation subsample must be used).

The results are reported in columns 7 and 8 of Tables 3, 4, and 5. The coefficient estimates for *PLANT PROPORTION* and *UTILITY PROPORTION* are very imprecise. They are of opposite signs and are neither individually nor jointly significant at conventional significance levels. The coefficient of *PLANT/UTILITY* is also very imprecise and varies in sign depending on estimating technique. Introducing these variables has no effect on the estimates for the primary variables of interest. These results imply that a plant or utility that relies on a single supplier for a large fraction of its requirements, or that depends heavily on a specific plant, does not encounter significant hold-up problems per se. The lock-in effect associated with designing plants to burn a particular type of coal becomes a potential contractual problem only to the extent that the other asset specificity characteristics are active.

[36] The inclusion of these three additional observations does not change the aggregate results.

[37] Including these variables in (1) and (3) does not change the results and I report the results for (2) in order to conserve space.

V. Contract Quantity

Before concluding, it is useful to explore the role of asset specificity in determining annual contract quantities since this variable plays such an important role in the contract duration equation. A relationship between contract quantity and asset specificity potentially emerges because of the presence of all three types of asset specificity. First, other things equal, physical asset specificity considerations suggest that a plant operator would like to rely on a specific supplier producing a particular type of coal at a particular location to the greatest extent possible.[38] This implies that contract quantity should vary directly with the coal requirements of the individual plant (*PLANT QUANTITY*).[39] On the other hand, the more a utility comes to rely on a single supplier the more costly a breach of contract may be. This suggests that a utility may be willing to rely more on a single supplier for an individual plant the larger is total utility utilization of coal (*UTILITY QUANTITY*) given the utilization of a specific plant.

Second, in the case of mine-mouth plants, the nature of the *ex ante* location/investment decision involves the mutual expectation that all or most of a plant's requirements will be taken from the proximate supplier. This implies that the quantity per contract will be larger for mine-mouth plants, other things equal.

Finally, as discussed above, when utilities design plants to closely match specific coal quality attributes they will have an interest in relying more heavily on a specific supplier who contracts to supply coal from a seam with these characteristics. This is likely to be an especially important consideration for coal supplies from the western region.

I estimate the following relationship using the single-plant (169 observation) subsample to determine empirically whether and how these considerations affect annual contract quantities.

$$\begin{aligned}(5)\quad QUANTITY_i = c_0 &+ d_1 PLANT\text{-}QUANTITY_i \\ &+ d_2 PLANT\text{-}QUANTITY_i^2 \\ &+ d_3 UTILITY\text{-}QUANTITY_i \\ &+ d_4 UTILITY\text{-}QUANTITY_i^2 \\ &+ d_5 MINE\text{-}MOUTH_i \\ &+ d_6 MIDWEST_i + d_7 WEST_i + v_i.\end{aligned}$$

I expect d_1, d_3, d_5, d_6, and d_7 to be greater than zero, and d_7 should be larger than d_6.

Equation (5) is estimated in two different ways. First, OLS estimates are presented. Second, OLS estimates with a correction to reflect the possibility that I have a censored sample are also presented. The OLS estimates of the coefficients of (5) may be biased as a result of the sampling procedure discussed earlier. We observe contract quantity only if contract duration is greater than or equal to (1979 − contract *YEAR*), so we have a censored sample. This implies that the random error (v) in the contract quantity equation (5) may be correlated with the random error (u) in the contract duration equation. If this is the case, the random error (v) in the contract quantity equation will be a function of the independent variables in the contract duration equation. The OLS estimates of the coefficients of the independent variables in the quantity equation (5) would then be biased if they are correlated with the independent variables in the duration equation. In particular, independent variables that appear in the contract duration equation may appear to be significant when introduced as independent variables in equation (8) when in fact they are not.[40] Since three

[38] One might also argue that, other things equal, a buyer would rather rely on a single supplier to conserve on more traditional types of transactions cost associated with negotiating, monitoring, and enforcing contracts with multiple supplies.

[39] Ideally, we would like to look at specific generating *units* rather than specific generating *plants* where plants have multiple units with different design characteristics. Unfortunately, coal supply information is not available at the generating unit level.

[40] See George Judge et al. (1985, pp. 610–13) and James Heckman (1976, 1979).

TABLE 7—CONTRACT QUANTITY[a]

	Dependent Variable: *QUANTITY*			
	OLS		OLS/*H*	
Independent Variables	(1)	(2)	(3)	(4)
PLANT QUANTITY	0.2501	0.3133	0.2128	0.2827
	(0.0441)	(0.1573)	0.0489	(0.1578)
PLANT QUANTITY-SQUARED	–	−0.00066	–	−0.00067
		(0.0010)		(0.00099)
UTILLITY QUANTITY	0.00349	0.0672	0.00393	0.06376
	(0.00689)	(0.0311)	(0.00685)	(0.03104)
UTILITY QUANTITY-SQUARED	–	−0.00007	–	−0.000063
		(0.00003)		(0.000032)
MINE-MOUTH	29.7932	27.1594	30.3261	27.775
	(6.8591)	(6.8997)	(6.8251)	(6.879)
WEST	13.5416	13.5865	8.4955	8.9738
	(4.6437)	(4.6259)	(5.4680)	(5.4597)
MIDWEST	5.2469	4.856	2.5832	2.4749
	(4.1841)	(4.1708)	(4.4371)	(4.4193)
H	–	–	−66.7618	−60.6647
			(38.7474)	(38.5749)
Constant	1.7329	−4.2283	10.6708	4.0997
	(3.5879)	(5.0515)	(6.2951)	(7.3027)
Corrected *R*-Squared	0.33	0.34	0.34	0.35
Observations	169	169	169	169

[a] Standard errors of coefficient estimates are shown in parentheses.

variables that appear in the quantity equation (5) also appear in the duration equation, this is a potential problem here.

We can obtain consistent estimates of the coefficients of (5) by obtaining maximum likelihood estimates of a reduced-form contract duration equation[41] to generate a sample selection correction H for each observation, adding the estimated values of H to (5) and then estimating the augmented equation (5) using OLS. The coefficient of H is then a consistent estimate of the covariance of u and v.

The results are reported in Table 7.[42] The OLS results appear in columns 1 and 2 and the OLS results with H introduced appear in columns 3 and 4. In both cases, estimates with and without quadratic terms in plant and utility quantities are reported. The results are generally consistent with my expectations. Larger plants tend to place larger orders and utilities with larger aggregate requirements do so as well, although the utility effect is generally small. The quadratic in plant quantity is not significantly different from zero. The coefficients of the mine-mouth dummy has the predicted sign and is estimated fairly precisely. Mine-mouth plants have contracts that are nearly 1.5 million tons larger (30 trillion Btu's) than other plants *ceteris paribus*. Regional differences in supply characteristics lead to larger contracts with western suppliers than with suppliers elsewhere. The coefficient of the correction variable H is negative, although not quite significant at the 5 percent level (two-tailed test), implying that the errors in the duration and quantity equation are negatively correlated. Including this correction does not have dramatic effects on the results for the coefficients of interest, however. The primary effect is to reduce the magnitude and significance of the coefficient of *WEST*.

[41] *QUANTITY* and *LOG-QUANTITY* are treated as being endogenous.

[42] The mean value for H is 0.076.

Finally, for the record, I report two-stage least square (*2SLS*) estimates of equation (2) in Table 3, column 9, and the equivalent of two-stage least squares for the maximum likelihood estimates of equation (2) in Table 5, column 9.[43] The estimates do not change in any important way from those obtained using the other estimating techniques in either case.

VI. Conclusions

The purpose of this paper has been to examine empirically hypotheses about the relationship between the duration of coal contracts and the presence of the three types of relationship specific investments discussed by Williamson (1983).[44] I argue that as relationship-specific investments become more important, the parties will find it advantageous to rely on longer-term contracts that specify the terms and conditions of repeated transactions *ex ante*, rather than relying on repeated bargaining. I make use of a large sample of coal contracts to examine this hypothesis. The empirical results obtained provide fairly strong support for this hypothesis. They are quite robust to alternative model specifications, samples, and estimating techniques. The results therefore provide additional empirical support for the view that the structure of vertical relationships between buyers and sellers is strongly affected by variations in the importance of relationship-specific investments.

APPENDIX

Here I discuss the sources of the data and the construction of the variables. The data base that I rely on was constructed from a variety of sources for use in a research project focusing on vertical relationships between electric utilities and coal suppliers. This is one of three papers that has been produced so far from this project.

The construction of the data base began with the choice of contracts to use in the analysis. Contracts were chosen if they appeared in *both* the 1981 and 1983 editions of *The Guide to Coal Contracts* (Pasha Publications) and for which information necessary for the project was reported. Contracts had to appear in both publications because some information that was desired appeared in one or the other publication, but not both. This also made it possible to check for errors and inconsistencies in the contract characteristics reported. To appear in both publications, contracts had to be in force in 1979. In five cases, actual contracts were used to supplement the data available from the primary sources.

This collection procedure resulted in a sample of 296 contracts (of which 277 had enough information to be used here) which generally had the following information, some of which is used in this paper and some of which I am using in related work:

1. Information required to calculate the agreed upon duration of the contract (see discussion below).
2. Contract quantities for 1979, 1980, and/or 1981 in tons.
3. The contract specifications for the Btu content and the sulfur content of the coal.
4. The identity of the seller and the location of the mine.
5. The identity of the buyer and the destination of the coal.
6. The base price for the coal at the time the contract was signed.
7. The actual price for the coal in 1979, 1980, and/or 1981.
8. Delivered quantities in 1979, 1980, and/or 1981.
9. Actual Btu and sulfur content of the coal.

Once the contract sample was selected, individual contracts were matched with specific utilities and power

[43] I assumed that *LOG-QUANTITY* is endogenous, and use the right-hand side variables in (5) as instruments for the *2SLS* results reported in Table 3, col. 9. Obtaining the equivalent of *2SLS* estimates for the case in which I treat the sample as being truncated and use maximum likelihood techniques, is more complicated. Following Maddala (pp. 234–40) and L. S. Lee et al. (1979), I proceeded in the following way. First, I estimate a reduced-form duration equation using maximum likelihood techniques. This allows me to obtain consistent estimates of H_i which can in turn be used to obtain consistent estimates of a contract quantity equation as discussed above. I use these estimates of the quantity equation to obtain predicted values of *QUANTITY* or *LOG-QUANTITY* which are then used in place of *QUANTITY* and *LOG-QUANTITY* to estimate equations (1) and (2) using maximum likelihood techniques. The results reported in Table 5, col. 9, assume that log(*PLANT QUANTITY*), log(*UTILITY QUANTITY*), *MIDWEST*, *WEST*, and *MINE-MOUTH* are exogenous variables. These variables are used to estimate a reduced-form duration equation using the maximum likelihood technique described earlier. An equation for log(*QUANTITY*) is then estimated using log(*PLANT QUANTITY*), log(*UTILITY QUANTITY*), *MIDWEST*, *WEST*, *MINE-MOUTH*, and *H* (generated from the reduced-form duration estimates) using OLS. The predicted values from this equation are then used instead of *LOG-QUANTITY* in (2) to obtain the maximum likelihood estimates reported. Specification (1) has also been estimated using this approach. The results appear to be robust.

[44] As well as similar considerations of transaction-specific investments identified by Klein et al.

plants (where possible). Two publications were utilized to obtain coal quantity and quality (Btu content) information by plant and utility: *Cost and Quality of Fuels for the Electric Utility Industry* (U.S. Department of Energy, various years) and *Steam Electric Plant Factors* (National Coal Association, various years). For public utility holding companies, coal utilization for subsidiaries was aggregated. Jointly owned plants were assigned to the operating company.

The three coal supply regions represent aggregations of smaller U.S. Bureau of Mines (BOM) districts. The West was defined as including BOM districts 16 through 23. The Midwest included BOM districts 5, 9, 10, 11, 12, 14, and 15. BOM district 15 includes Texas, but we have no contracts for coal produced in Texas. The East includes coal from the remaining BOM districts, primarily in Appalachia. The differences between regions is discussed in more detail in my 1985 paper. The *Keystone Coal Industry Manual* (Mining Information Services) and an atlas were used to help locate mines in specific BOM districts.

The variable definitions and construction are as follows:

DURATION: Contract Duration: The contract data base generally provides information for the date the contract was executed, the termination date and (less frequently) the date of first delivery of coal. A specific month and year is often provided, but sometimes the source specifies only years. Because the contract execution date was available more often than the date of first delivery and because the two are generally quite close together, contract duration was measured as contract termination year minus execution year. Initial experimentation with duration measured using the date of first delivery or using month and year indicated that the results were unaffected, so the definition that preserved the largest number of contracts was used.

QUANTITY: Annual contract quantity in trillion Btu's. The contracted tonnage reported for 1980 (if that was not available, or obviously not representative, 1979 or, alternatively, 1981 were used instead) was multiplied by the contracted Btu content of the coal to arrive at the contract quantity variable.

MINE-MOUTH: Mine-mouth dummy variable that is equal to one if the plant is a mine-mouth plant and zero otherwise. The information in my 1985 paper combined with the coal destination information in the *Guide To Coal Contracts* was used to construct this variable. The Navajo and Mohave plants were included in this category as well since they have economic characteristics very much like a mine-mouth plant.

MIDWEST: A regional dummy variable that equals one if the coal is from a midwestern mine (as defined above) and zero otherwise. The contract data base provides information on mine location.

WEST: A regional dummy variable that equals one if coal is from a western mine (as defined above) and zero otherwise.

PLANT QUANTITY: Annual plant utilization of coal. Coal utilization by a plant to which a specific contract is dedicated (at least 90 percent of the coal delivered to a single plant) for 1980 (1979 or 1981 if necessary to match *QUANTITY*) in trillion Btu's. Obtained from the Department of Energy and National Coal Association publications identified above.

UTILITY QUANTITY: Annual utility utilization of coal. Coal utilization in 1980 (or 1979 or 1981 if necessary to match other data) by the utility operating a plant to which a contract is dedicated. Obtained from the Department of Energy and National Coal Association publications identified above.

PLANT PROPORTION: Delivered contract quantity in Btu's divided by plant utilization in Btu's. Delivered contract quantity in tons was pulled from the contract information for 1980 (or 1979 if 1980 was not available), and multiplied by the delivered Btu content of the coal to obtain quantities delivered to a specific plant under a contract dedicated to that plant. This figure was then divided by *PLANT QUANTITY*.

UTILITY PROPORTION: Delivered contract quantity in Btu's (as defined above in definition of *PLANT PROPORTION*) divided by utility utilization of coal in Btu's for 1980 (1979 or 1981 otherwise).

PLANT/UTILITY: *PLANT QUANTITY* divided by *UTILITY QUANTITY*.

YEAR: The year specified as the execution date of the contract.

DATE-71: A dummy variable that equals one for contracts signed in 1971, 1972, and 1973.

DATE-74: A dummy variable that equals one for contracts signed in 1974, 1975, 1976, and 1977.

DATE-78: A dummy variable that equals one for contracts signed in 1978 and 1979.

The mean, standard deviation, minimum, and maximum values of these variables for the 277 observation sample and the 169 observation (single delivery point) subsample are contained in Tables 1 and 2. The data for the variables used in this paper are available upon request.

REFERENCES

Crocker, Keith J. and Masten, Scott E., "Mitigating Contractual Hazards: Unilateral Options and Contract Length," Working Paper No. 449, Graduate School of Business Administration, University of Michigan, March 1986.

Dubin, Jeffrey and Rivers, R. Douglas, *Statistical Software Tools*, Version 1.0, Pasadena, March 1986.

Goldberg, Victor and Erickson, John, "Long Term Contracts for Petroleum Coke," Working Paper No. 206, Department of Economics, University of California-Davis, 1982.

Hart, Oliver and Holmstrom, Bengt, "The Theory of Contracts," Department of Economics Working Paper No. 418, MIT, March 1986.

Heckman, James, "The Common Structure of

Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator For Such Models," *Annals of Economic and Social Measurement*, 1976, Vol. 5, 475–92.

______ "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, *47*, 153–61.

Joskow, Paul L., "Vertical Integration and Long Term Contracts: The Case of Coal-burning Electric Generating Plants," *Journal of Law, Economics and Organization*, Spring 1985, *1*, 33–80.

______ "Price Adjustment in Longer Term Contracts: The Case of Coal," mimeo., May 1986.

______ **and Mishkin, Frederick,** "Electric Utility Fuel Choice Behavior in the United States," *International Economic Review*, October 1977, *18*, 719–36.

______ **and Schmalensee, Richard,** "The Performance of Steam Electric Generating Plants in the United States: 1960–1989," Department of Economics Working Paper No. 379, MIT, July 1985.

Judge, George G. et al., *The Theory and Practice of Econometrics*, New York: Wiley & Sons, 1985.

Klein, Benjamin, Crawford, Robert and Alchian, Armen, "Vertical Integration, Appropriable Rents and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Lee, L. S., Maddala, G. S. and Trost, R. P., "Testing for Structural Change By D-Methods in Switching Simultaneous Equations Models," *Proceedings of the American Statistical Association*, Business and Economics Section, 1979, 461–66.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Masten, Scott E., "The Organization of Production: Evidence from the Aerospace Industry," *Journal of Law and Economics*, October 1984, *27*, 403–18.

______ **and Crocker, Keith, J.,** "Efficient Adaptation in Long-Term Contracts: Take-or-Pay Provisions for Natural Gas," *American Economic Review*, December 1985, *75*, 1083–93.

Monteverde, Kirk and Teece, David, "Supplier Switching Costs and Vertical Integration in the Automobile Industry," *Bell Journal of Economics*, Spring 1982, *13*, 206–13.

Mulhern, J. Harold, "Complexity in Long-term Contracts: An Analysis of Natural Gas Contract Provisions," *Journal of Law, Economics and Organization*, Spring 1986, *2*, 105–117.

Schmalensee, Richard and Joskow, Paul L., "Estimated Parameters as Independent Variables: An Application to the Costs of Steam Electric Generating Units," *Journal of Econometrics*, April 1986, *31*, 275–305.

Williamson, Oliver E., "Transaction-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, *22*, 233–61.

______, "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, *73*, 519–40.

______, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Zimmerman, Martin, *The U.S. Coal Industry: Economics of Policy Choice*, Cambridge: MIT Press, 1981.

Mining Information Services, *Keystone Coal Industry Manual*, New York, 1981.

National Coal Association, *Steam Electric Plant Factors*, Washington, various years.

Pasha Publications, *Guide To Coal Contracts*, Washington, 1981; 1983.

U.S. Department of Energy, Energy Information Administration, *Cost and Quality of Fuels for Electric Utility Plants*, DOE/IEA-0191, Washington, various years.

Diamonds are a Government's Best Friend: Burden-Free Taxes on Goods Valued for their Values

By YEW-KWANG NG*

To most economists, it seems almost axiomatic that taxes (except corrective taxes) impose not just a burden equal to the amount of the tax collected, but also an excess burden by distorting individual choices, not to mention administrative, compliance, and policing costs (loosely called transaction costs below). Lump sum taxes with no excess burden exist only in theory. Yet there exists at least one class of goods that can be taxed and the tax will not only not produce an excess burden, but it will not be a burden at all (ignoring transaction costs)! This sounds like a miracle but it is really quite simple once it is recognized that some goods are valued for their values, not for their intrinsic consumption effects. Taxes on these goods increase their prices. But consumers can reduce quantities consumed without changing the values of these goods, suffering no loss in utility. Thus, no burden is imposed, not to mention excess burden. For example, after a doubling in the price of diamond, a $1000 gift of diamond is still valued at $1000, though the size of the stone is smaller.

A carat of diamond can be worth thousands of dollars, but costume jewelry that looks similar may cost only a few dollars. Imitation diamonds look virtually the same as real diamonds and it takes experts with fine instruments to tell the difference. Surely, diamonds are valued not for their intrinsic consumption effects but because they are costly. Consumers of diamonds either derive utility by showing off their wealth (Veblen's conspicuous consumption), by using it as a store of value, or by giving it as a gift of value. In virtually all cases, it is the value, not the diamond itself, that counts. This applies to most other precious stones and metals, including gold. In various lesser degrees, this "diamond effect" also applies to other items of conspicuous consumption such as expensive fur coats and luxurious cars. With increasing affluence, "diamond goods" will become more important.

The diamond effect (valuing something for its value rather than the consumption effect) must be distinguished from other similar or related but different phenomena. First, "the habit of judging quality by price" (Tibor Scitovsky, 1945) is the belief that higher-priced brands give higher intrinsic consumption utilities. Second, while Thorstein Veblen's (1899) conspicuous consumption may partly contribute to the diamond effect, they are conceptually distinct. A person's desire to go on a world trip may be partly to show off to his (her) admiring friends who cannot afford to go. But as long as he can and they cannot, it has value for conspicuous consumption. Increases in the price of the trip may add little to the value and half a world trip is not as good, even if the price of the world trip has doubled. On the other hand, a man's gift of a $1000 diamond ring to his wife is worth that much irrespective of the size of the stone, and they may never show the ring to anybody. Also, some people show off their heavy gold bracelets (conspicuous consumption with a diamond effect) while others hide their gold bullion (a diamond effect without conspicuous consumption). Third, there are goods whose intrinsic consumption effects depend on whether other people also consume them (for example, telephones, unusual clothing, fashions). To some extent, this consideration may also affect the diamond effect. However, to concentrate on the pure diamond effect, I will abstract these complications away.

As in the distinction between private and public goods where different degrees of pub-

*Department of Economics, Monash University, Clayton, Victoria 3168 Australia, and University of Maryland-College Park.

licness are involved, most goods are valued partly for their intrinsic consumption effects (approaching 100 percent for ordinary items like bread) and their values (approaching 100 percent for precious jewels). However, for analytical simplicity, I consider only two polar cases and adopt an atemporal model commonly used in welfare analysis. The complicated questions of the dynamics of transition and some other practical complications are touched on in the concluding section, but are not formally analyzed.

The basic result is that a change in the price of a diamond good leaves its own value unchanged and the amount of all other goods consumed, and hence the utility levels of consumers unchanged, and that it is optimal to place arbitrarily high taxes on diamond goods which impose no burden and no excess burden. A corollary is that the demand curve for a diamond good is a rectangular hyperbola with unit price elasticity throughout.

Such an obvious phenomenon as the diamond effect has not of course completely escaped the economist's attention. For example, Pigou touched on the "desire to possess what other people do not possess" (1932, p. 226) and used diamonds as an example. But there is a curious lack of formal analysis,[1] and almost complete disregard in the public finance texts (for example, Richard Musgrave and Peggy Musgrave, 1980) and actual policy debate on taxation issues (for example, the great Australian tax reform in 1985). This paper provides a modest attempt at a formal analysis which may attract, hopefully, more attention both by theorists and those concerned with policy decision.

I. A Simple Analysis

For simplicity, consider the case with only one diamond good (say, the first) and ignore all other complications (externalities, etc.). Generalization to impure diamond goods is straightforward, but impure or mixed goods bring some complications not considered in this paper. The utility function of an individual may thus be written as

$$(1) \qquad U\left(p^1x^1/p^n, x^2,\ldots,x^n\right),$$

where x^i is the amount of the ith good consumed, p^i its price, and the last good is being used as a numeraire. In the long run, the consumer is unlikely to suffer from significant money illusion. Thus, instead of the money value of the diamond good p^1x^1, we should replace the money price p^1 by real or relative price p^1/p^n.[2] For impure diamond goods, both p^1x^1/p^n and x^1 enter the utility function.

Taking prices as given, the consumer maximizes (1) subject to

$$(2) \qquad \sum p^ix^i = M,$$

where the summation is over all the n goods and M is the given amount of income. This problem is homogeneous of degree zero in all prices and money income; no money illusion is involved.

Assuming an interior solution for notational simplicity, the first-order conditions for optimality are

$$(3a) \qquad U_1 = \lambda p^n$$

$$(3b) \qquad U_i = \lambda p^i \qquad (i = 2,\ldots,n)$$

where U_i is the partial derivative (marginal utility) of the ith element in the utility function (1), and λ is the Lagrangian multiplier

[1] Peter Kalman (1968) provides a rigorous analysis of consumer behavior when prices enter the utility function. This is a very general analysis which may be said to include "judging quality by price," conspicuous consumption, and the diamond effect. However, partly because it is too general and partly because of its exclusive concern with the positive theory of consumer behavior, it reaches none of the results of this paper.

[2] Alternatively, we may replace p^n in (1) by P, a price index of all nondiamond goods, with the same result except that p^n below is replaced by P.

associated with (2) or the marginal utility of income.

The price of a diamond, p^1, does not appear in the system of equations (3) describing the optimal solution. It is tempting but wrong to infer from this that the optimal x's are independent of p^1. This is wrong because p^1 appears in (2) which is included in the set of equations, together with (3), defining the optimal solution. However, it is true that p^1x^1/p^n and $x^2,\dots,x^n$ and hence U are independent ofp^1, as shown below.

The maximization problem above may be written, with no change of any substance, as the maximization of

$$U(y^1, y^2,\dots, y^n) \tag{4}$$

subject to

$$\sum q^i y^i = M, \tag{5}$$

where

$$y^1 \equiv p^1x^1/p^n, \quad y^i \equiv x^i \quad (i=2,\dots,n),$$
$$q^1 \equiv p^n, \quad q^i \equiv p^i \quad (i=2,\dots,n).$$

This rewritten problem is identical in its mathematical form to the traditional consumer optimization problem with no diamond effect, and with the following familiar first-order conditions for an interior solution,

$$U_i = \lambda q^i \quad (i=1,\dots,n), \tag{6}$$

which, with constraint equation (5), define the optimal y's.

With the rewritten problem, if we work in terms of y's instead of x's (the only difference is to take p^1x^1/p^n as an integral variable instead of breaking it up into its constituent parts), it is clear that p^1 appears neither in the constraint (5) nor in the first-order condition (6). The optimal set of y's and hence the maximized utility level are thus independent of p^1. This result may be expressed as

PROPOSITION 1: *A change in the price of a diamond good leaves its value and the amounts of all other goods consumed, and hence the utility level of the consumer unaffected.*

COROLLARY 1: *The demand curve for a diamond good is a rectangular hyperbola with unit elasticity throughout the whole range where it remains a pure diamond good.*

This is true not only for an individual demand curve, but also for a market demand curve as long as the good is viewed by all consumers as a diamond good (assumed here for simplicity) because the horizontal summation of rectangular hyperbolas is also a rectangular hyperbola.

For simplicity, assume a horizontal supply curve. A 100 percent tax on the diamond good then doubles its price and a 200 percent tax triples it, etc. The higher the tax rate, the larger the tax revenue, while consumers remain no worse off. The tax revenue collected thus represents pure gain, imposing not only no excess burden, but also no burden at all!

There is an upper limit beyond which the tax revenue cannot exceed. This supremum (the maximum does not exist) is the pre-tax (= post-tax) value of the good. The amount of tax revenue that can be raised without burden is limited by the amount of expenditure on diamond goods (which may be expected to increase relatively and absolutely with increasing affluence). The net gain is the amount of resources saved due to a smaller output after the imposition of the diamond tax.

II. A Model of Optimal Taxation

The above analysis may be regarded as somewhat partial and/or intuitive. Here, I present a standard model of optimal taxation, except that I allow for pure diamond goods. Since no changes in the relative price between private goods is considered, I lump them into a composite good y. Similarly, I lump all diamond goods into another composite good d. As in the standard optimal taxation literature, I concentrate on the tax side by assuming a constant government revenue requirement and assume that the consumer side of the economy can be represented by one consumer or a community utility function,

$$U(D, y), \tag{7}$$

where $D \equiv (q+t)d/(Q+T)$ is the (relative) value of the diamond good, q and Q are the fixed producer prices, and t and T the per unit taxes on diamond and the private good, respectively. The assumption of a representative consumer does assume away distributional issues, but may be justified by the predominant concern on efficiency issues and the argument on separating equity and efficiency issues even in the presence of second-best factors and other complications (see my 1984 paper).

The consumer maximizes (7) with respect to d and y, subject to

$$(q+t)d + (Q+T)y = M, \tag{8}$$

where money income M is taken as given. While this may seem to abstract away work-leisure choice, we may alternatively interpret M as full income and include leisure in the composite good y, with the result that leisure is regarded as taxable. If I can establish the result on the optimality of imposing a high tax on diamond even in a model where leisure is taxable, the desirability of doing so where leisure is not taxable seems to apply a fortiori.

The first-order condition for the consumer maximization is

$$U_D/U_y = 1, \tag{9}$$

where U_D is the marginal utility of the value of diamond (relative to the price of private good) consumer and U_y the marginal utility of the private good. In other words,

PROPOSITION 2: *In equilibrium, the marginal rate of substitution between the (relative) value of diamond and the private good equals unity.*

This may appear too simple to be true. But if I write the budget constraint (8) as

$$(Q+T)D + (Q+T)y = M, \tag{8$'$}$$

it can immediately be seen that the consumer price $(Q+T)$ of the private good serves as the price for both the private good y and for the relative value of diamond D, and equation (9) is thus obvious. The consumer allocates his (her) income M between two goods D and y that have the same price, so $MRS = 1$ for an interior maximum.

As discussed in Section I, the consumer's optimal choice between D and y is independent of the consumer diamond price, $q+t$, which entered neither (8′) nor (9). We thus have

$$\partial D/\partial t = 0 = \partial y/\partial t. \tag{10}$$

From the first inequality in (10) and the definition of D, we have

$$\eta^{dt} = -t/(q+t), \tag{11}$$

where $\eta^{dt} \equiv (\partial d/\partial t)d/t$ is the elasticity of d with respect to t.

The government maximizes (7) with respect to t and T, subject to the consumer's choice described above and to the fixed revenue constraint

$$dt + yT = \bar{R}. \tag{12}$$

The first-order conditions for an interior solution are

$$\left(\frac{d}{Q+T} + \frac{q+t}{Q+T}\frac{\partial d}{\partial t}\right)U_D + \frac{\partial y}{\partial t}U_y = \theta\left(d + t\frac{\partial d}{\partial t} + T\frac{\partial y}{\partial t}\right) \tag{13}$$

$$\left\{\frac{q+t}{Q+T}\frac{\partial d}{\partial T} - \frac{(q+t)d}{(Q+T)^2}\right\}U_D + \frac{\partial y}{\partial T}U_y = \theta\left(y + T\frac{\partial y}{\partial T} + t\frac{\partial d}{\partial T}\right), \tag{14}$$

where θ is the Lagrangian multiplier associated with (12). Eliminate θ between (13) and (14) and rewrite expressions in elasticity form, that is, $\eta^{xy} = (\partial x/\partial y)y/x$, we have

(15)

$$\frac{\frac{d}{t}\left(\frac{t}{Q+T}+\frac{q+t}{Q+T}\eta^{dt}\right)U_D+\frac{y}{t}\eta^{yt}U_y}{\frac{d}{T}\left\{\frac{q+t}{Q+T}\eta^{dT}-\frac{(q+t)T}{(Q+T)^2}\right\}U_D+\frac{y}{T}\eta^{yt}U_y}$$

$$=\frac{d(1+\eta^{dt})+\frac{yT}{t}\eta^{yt}}{y(1+\eta^{yT})+\frac{dt}{T}\eta^{dT}}.$$

Substitute η^{dt} from (11) and $\eta^{yt} = 0$ (from the second equality in equation (10)) into (15), the numerator of the left-hand side of (15) becomes zero. The denominator of the right-hand side does not equal infinity unless t itself is infinite, since an infinitesimal change in T (equivalent to a reverse change in M) does not cause a jump in either y or d under traditional assumptions about the consumer. Therefore, the numerator of the right-hand side must equal zero. Since $\eta^{yt} = 0$ (from equation (10)), $d \neq 0$ for an interior solution, we have $1 + \eta^{dt} = 1 - t/(q+t)$ (from equation (11)) $= 0$), or

$$(16) \qquad t/(q+t) = 1.$$

Since q is nonzero, (16) can hold if and only if t is infinite. This gives us

PROPOSITION 3: *A pure diamond good has an infinite tax in an optimal tax system.*

This result confirms the analysis of Section I. Of course, in practice, as the tax on diamonds gets to be very high, the physical amount of diamonds of a given value becomes very small. This will eventually affect the intrinsic consumption value of diamonds, or at least increase the cost of handling tiny quantities. Thus my model of pure diamond goods ceases to be an accurate approximation when t becomes very high. Taking account of this, a very high tax rather than an infinite tax is optimal.

III. Concluding Remarks

Other practical considerations also suggest a reasonably high, instead of an arbitrarily huge, tax on diamond goods. Too high a tax induces tax evasion (including smuggling), especially if only a few countries are imposing high taxes. This suggests that international cooperation to raise taxes on diamond goods may be desirable.

My analysis, being conducted in an atemporal framework, also ignores the complication of dynamic transition. Ideally, when taxes on diamond goods are introduced, the preexisting possessors of diamond goods should also pay the taxes. This enlarges the tax base (and hence the tax revenue) and also avoids the distributional problem if only new diamond goods are taxed. This problem arises because existing possessors of diamond goods are actually made better off by the taxes, while new consumers of diamond goods are not. However, it may not be administratively and politically feasible to tax existing diamond goods. The distributional problem may then suggest that taxes on diamond goods should be lower. The following example will make clear this distributional-dynamic transition issue. Assume that a 100 percent tax on all future production (assumed competitive) or consumption of a (set of) diamond good is imposed. A distributionally neutral policy is to impose the same 100 percent tax on existing stocks in the form of conscripting 50 percent of all holdings which are then destroyed. The gain in revenue consists only in taxes on future production. No one is worse off as all diamond goods double in prices. If the conscripted 50 percent are not destroyed but put back into the market (as equivalent, ignoring transaction costs, to a 100 percent monetary tax on existing stocks), this will depress the prices of diamond goods (from their doubled values) and make existing holders worse off. However, reasonably assuming continuity, there exists a tax rate t $(0 < t < 100\%)$ on existing holdings that will leave them indifferent. The government will then immediately gain the revenue from taxes on existing holdings, at a cost (in comparison to the alternative of destroying half of the existing holdings) of foregoing revenue in the immediate future when no diamond goods are produced, at least not from existing marginal producers, since prices are below marginal production costs plus taxes. The transition raises some interesting dynamic problems (including the relative desirability of alternative policies). However, a detailed analysis requires an explicitly dynamic model beyond the scope of this paper.

In any case, the argument for treating a dollar as a dollar to whomsoever it goes (see my earlier paper) suggests that it is better to impose the full burden-free taxes on efficiency considerations with possible adjustments in income and wealth taxes to achieve the objective of equality.

REFERENCES

Kalman, Peter J., "Theory of Consumer Behavior when Prices Enter the Utility Function," *Econometrica*, July-October 1968, *36*, 497–510.

Musgrave, Richard A. and Musgrave, Peggy B., *Public Finance in Theory and Practice*, 3rd ed., New York: McGraw-Hill, 1980.

Ng, Yew-Kwang, "Quasi-Pareto Social Improvements," *American Economic Review*, December 1984, *74*, 1033–50.

Pigou, Arthur C., *The Economics of Welfare*, 4th ed., London: Macmillan, 1932.

Scitovsky, Tibor, "Some Consequences of the Habit of Judging Quality by Price," *Review of Economic Studies*, No. 2, 1945, *12*, 100–05.

Veblen, Thorstein, *The Theory of the Leisure Class*, New York: Macmillan, 1899.

A Note on Indivisibilities, Specialization, and Economies of Scale

By BRIAN K. EDWARDS AND ROSS M. STARR*

It is well known that factor indivisibilities and opportunities for labor specialization (division of labor) can result in scale economies. We argue that the second observation is a special case of the first; labor specialization results in scale economies only through indivisibility or other nonconvexity in the use of labor. Hence, the observation that labor specialization results in scale economies is correct but a half-truth; it relies on the unstated assumption of indivisibility or nonconvexity in the use of labor.[1] Tjalling Koopmans (1957) citing E. H. Chamberlin (1948) and Nicholas Kaldor (1934), described this observation,

> The relevant aspect of worker specialization appears to be that, up to a certain degree of specialization, the undivided attention given by a specialized worker to a full-time task of a sufficiently challenging character produces not exactly (but presumably more than) twice as much as half-time attention (with half the training!) given to the same task, if the other half of the worker's time (and training) is applied to a different productive activity.
> [p. 151]

Adam Smith noted that "division of labor is limited by the extent of the market." On the contrary, if labor were fully divisible, Smith's statement would be false; there would be no particular reason why market size should pose a limitation on division of production tasks. If, however, labor is indivisible or displays nonconvexity in use, then Smith's statement is correct. Sufficient scale would be required to overcome indivisibilities to allow (indivisible) labor to specialize in separate portions of the production process. Alternatively, a setup cost (a nonconvexity) in the transition of labor between production operations is a sufficient condition for scale to be required to reduce average cost, through reduction of frequency of switching operations. Finally, if a setup cost in training time is needed for acquisition of a specialized skill, this nonconvexity will account for a scale economy in the employment of specialized labor for the production sector (but not necessarily for the individual firm).

I. The Pin Factor Example

In his pin factory analysis, Smith recognized the role of nonconvexity in labor use, attributing much of "the great increase in the quantity of work... in consequence of the division of labor... to the saving of the time which is commonly lost in passing from one specie of work to another" (p. 7). Hence, in Smith's view, employing the same worker at different tasks requires incurring a transition setup cost. Given sufficient scale, it is preferable to allow labor to specialize and avoid this switching cost, that is, to use labor in indivisible increments.

The production of pins will involve choosing one of many possible techniques of production. The crudest technology will involve using the same worker in all operations of

*Staff Economist, Office of the Chief Economist, United States General Accounting Office, 441 G Street, NW, Rm 4001, Washington, D.C. 20548, and Professor of Economics, Department of Economics, D-008, University of California-San Diego, La Jolla, CA 92093. The views expressed in this paper are our own, and do not represent those of the United States General Accounting Office.

[1] Indivisibilities need not be the sole rationale for scale economies in capital. Nicholas Kaldor (1972) attributes scale economies in part to the "three-dimensional nature of space." That is, that production capacity in some processes will vary with physical volume of plant or equipment while cost may depend principally on surface area, the latter varying as the two-thirds power of volume.

production. More sophisticated means of production are defined by having each unit of labor assigned to fewer operations.

> One man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make a head requires two or three distinct operations: to put it on, is a peculiar business, to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufactories, are all performed by distinct hands, though in others the same man will sometimes perform two or three of them. [Smith, pp. 4–5]

As an illustration of this example, consider a family of production functions, indexed by $k=1,\dots,n$, by which a single output, y, is produced. Although actual production of y will involve only one of the functions, progressively higher output levels will be achievable for a given set of inputs by using a more specialized production function. The limitation on specialization will be indivisibility or other nonconvexity, so that higher levels of specialization will be available only with sufficient input units. Consider the primary production function, defined by

$$(1)\qquad y=f_k(x_1,\dots,x_k)=b_k\prod_{i=1}^{k}x_i^{\alpha_{ik}}$$

where $y=$ output under process $k=1,\dots,n$; where $x_i=$ quantity of labor pursuing specialty i, $i=1,\dots,k$; for each k, $\Sigma_{i=1}^{k}\alpha_{ik}=1$, $\alpha_{ik}\geq 0$; where $b_k=$ technology parameter for process k; and $b_{k+1}>((k+1)/k)b_k$.

We have J workers, $j=1,\dots,J$. The variable x_{ij} is the amount of labor in specialty i provided by worker j: $x_i=\Sigma_{j=1}^{J}x_{ij}$. According to (1) there are n possible separate production processes, ranging from the simplest involving production by using only one task, ($k=1$, i.e., one class of labor input), to more complex ones involving many.

To convert f_k from constant returns to increasing returns let

$$(2)\qquad F_k(x_{11},\dots,x_{ij},\dots,x_{kJ})=f_k\left(\sum_{j=1}^{J}[x_{1j}],\dots,\sum_{j=1}^{J}[x_{kj}]\right),$$

where $[x_{ij}]$ denotes the greatest integer $\leq x_{ij}$; or let

$$(3)\qquad G_k(x_{11},\dots,x_{ij},\dots,x_{kJ})=f_k(x_1,\dots,x_k)-\sum_{i=1}^{k}\sum_{j=1}^{J}c_i\{x_{ij}\},$$

where $c_i>0$ and $\{x_{ij}\}$ is defined to be

$$0\quad\text{for } x_{ij}=0\qquad\text{and}\qquad 1\quad\text{for } x_{ij}>0.$$

To represent the sources of scale economy, we consider indivisibility (2), and setup cost (3). The advantages of specialization are embodied in the assumption that $b_{j+1}>((j+1)/j)b_j$. As a result, the production functions defined in equations (2) and (3) have the following characteristics: 1) for a given value of labor input, $\Sigma_i\Sigma_jx_{ij}$, production has higher average product (value of f_k) as we move from lower-order to higher-order processes within limits imposed by indivisibility and setup cost; 2) higher-order processes involve a finer division of inputs; and 3) while production within each process is characterized by constant returns to scale, the presence of indivisibilities or setup costs in the use of inputs results (though not uniformly) in increasing returns to scale.

II. Labor Specialization and Scale Economies in the Elementary Literature

Indivisibilities in equipment and the desirability of labor specialization are cited in the elementary literature as distinct sources of scale economies in production. Indivisibility (or other nonconvexity) of labor is seldom made explicit. Edwin Mansfield's text (1976) is typical of treatments that use labor specialization as a rationale for scale economies without treating indivisibility of labor as a logical step:[2]

> Some inputs are not available in small units; for example, we cannot install half an open hearth furnace. Because of indivisibilities of this sort, increasing returns to scale may occur.... Greater specialization also can result in increasing returns to scale; as more men and machines are used, it is possible to subdivide tasks and allow various inputs to specialize.
>
> [Mansfield, pp. 128, 129]

III. Conclusion

The possibility of productively superior specialization of labor is not, in itself, a sufficient condition for the presence of scale economies in production. The link between specialization (division of labor) and scale economies is indivisibility or other nonconvexity in application of labor. The classic treatment of Smith implicitly recognized this point and it was elaborated by Koopmans. Nevertheless, it is not explicit in the current elementary literature, which thereby obscures the logic of the analysis.

[2] The following standard texts recognize scale economies without going into detail as to their relationship to labor specialization: Charles Baird (1975), James Gwartney and Richard Stroup (1980), David Kamerschen and Lloyd Valentine (1981), E. Warren Shows and Robert Burton (1972). Alternatively, Stanley Kaish (1976), Paul Samuelson (1980), and Donald Watson and Malcolm Getz (1981) treat specialization and scale in a fashion similar to Mansfield. Richard Lipsey and Peter Steiner (1975) is an exception in explicitly recognizing indivisibility in use of labor. Joan Robinson and John Eatwell (1973) emphasize organizational and distributional issues rather than technology in their discussion of division of labor.

REFERENCES

Baird, Charles W., *Prices and Markets: Micròeconomics*, St. Paul: West Publishing, 1975.

Chamberlin, E. H. *The Theory of Monopolistic Competition*, 6th ed., Cambridge: Harvard University Press, 1948.

Gwartney, James D. and Stroup, Richard, *Microeconomics: Private and Public Choice*, New York: Academic Press, 1980.

Kaish, Stanley, *Microeconomics: Logic, Tools, and Analysis*, New York: Harper and Row, 1976.

Kaldor, Nicholas, "The Equilibrium of the Firm," *Economic Journal*, March 1934, *44*, 61–76.

______, "The Irrelevance of Equilibrium Economics," *Economic Journal*, December 1972, *82*, 1237–55.

Kamerschen, David R. and Valentine, Lloyd M., *Intermediate Microeconomic Theory*, 2nd ed., Cincinnati: South-Western, 1981.

Koopmans, T. C., *Three Essays on the State of Economic Science*, New York: McGraw-Hill, 1957.

Lipsey, Richard G. and Steiner, Peter O., *Economics*, 4th ed., New York: Harper and Row, 1975.

Mansfield, Edwin, *Microeconomics—Theory and Applications* (Shorter 2nd ed.), New York: W. W. Norton, 1976.

Robinson, Joan and Eatwell, John, *An Introduction to Modern Economics*, New York: McGraw-Hill, 1973.

Samuelson, Paul A., *Economics*, 11th ed., New York: McGraw-Hill, 1980.

Shows, E. Warren and Burton, Robert H., *Microeconomics*, Lexington: D. C. Heath, 1972.

Smith, Adam, *The Wealth of Nations*, New York: Modern Library, 1937.

Watson, Donald S. and Getz, Malcolm, *Price Theory and Its Uses*, 5th ed., Boston: Houghton Mifflin, 1981.

Rigidity vs. License

By JOSEPH FARRELL*

A central problem in the organization of society is to choose rules for making choices when people's interests conflict. Economists have had much to say about such problems when compensating payments are possible. In practice, however, such payments are often not made, and decisions are either rigidly imposed by a central authority or taken unilaterally by one of the parties concerned, who we say has "the right" to choose the outcome. In this paper, without asking why these rules are common, I ask which is most efficient.

I assume that the interested parties know more about their preferences than does the central authority (the "government"). This is a reason to give one of them the right to decide the outcome: in this model, rights are a decentralization device (see Partha Dasgupta, 1980), and since I assume a benevolent government, there is no role for decentralization without private information. But, while this private information *can* make the delegated decision more efficient than an undelegated and rigid central decision, there is a countervailing problem: an interested party chooses selfishly and ignores even what is common knowledge about others' preferences, which the benevolent though ignorant government would take into account. Thus we find that giving an interested party the right to choose is more desirable when he has important private information, but less desirable when he and others are very much in conflict.

I. A Simple Model of Conflict

Two agents, A and B, care about the choice of a real variable x. Their preferences also depend on private information: A knows the value of a random variable a which affects his payoff, while B knows the value of another (possibly correlated) random variable b, which affects hers. It is assumed in particular that payoffs are given by

$$u^A \equiv -\alpha(x-a)^2, \quad u^B \equiv -\beta(x-b)^2.$$

Thus a and b respectively represent A's and B's preferred outcomes. I assume that $a<b$ with probability one.[1] Neither a nor b is observable to a central planner. The parameters α and β represent the relative importance of the problem to A and B: I normalize by setting $\alpha+\beta=1$.

Write x^* for the optimal choice of x (given a and b) and x^n for the *ex ante* optimal choice of x. Since x^* maximizes $W \equiv u^A + u^B$, the first-order condition gives

$$2\alpha(x^*-a)+2\beta(x^*-b)=0,$$

whence

$$(1) \qquad x^*(a,b)=\alpha a+\beta b.$$

The *ex ante* optimum, x^n, is given by

$$(2) \qquad x^n=\alpha Ea+\beta Eb.$$

Evidently, x^n balances A's and B's conflicting preferences perfectly if the random variables a and b happen to take their expected values, but fails to respond to variations in a and b. Thus it represents the extreme of ignorant benevolence.

II. Efficiency Comparisons

I now compare the efficiency of three simple rules: "rights to A," in which A chooses

*University of California, Berkeley, CA 94720 and GTE Laboratories, Waltham, MA 02254. This paper is dedicated to the memory of M. J. Farrell (1926–75), who thought and cared about rights and efficiency.

[1] For instance, suppose that the support of a lies wholly below the support of b.

x; "rights to B," in which B chooses x; and "rigid norm," in which the choice $x = x^n$ is mandated.

I show that, if $\alpha \simeq \beta$ (the problem is about equally important to both parties), then the rigid norm is the most efficient rule unless there is "enough" correlation between a and b. However, if $\alpha \gg \beta$, then A should have the right to choose x.

If A has the right to choose x, he will set $x = a$, giving payoffs $u^A = 0$ and $u^B = -\beta(a-b)^2$. The *ex ante* expected value of welfare W is then

$$(3)\qquad EW(R^A) = -\beta E\{(a-b)^2\}.$$

Similarly, if B chooses x, she will set $x = b$, so that

$$(4)\qquad EW(R^B) = -\alpha E\{(a-b)^2\}.$$

From (3) and (4) we have

PROPOSITION 1: *It is more efficient to assign rights to the agent with more at stake than to the other agent.*

Finally, the rigid norm rule $R^n(x \equiv x^n)$ yields expected welfare:

$$(5)\quad EW(R^n) = -\alpha E\big((x^n - a)^2\big) - \beta E\big((x^n - b)^2\big),$$

which can be expanded to give

$$(6)\quad EW(R^n) = -\alpha\beta C^2 - \alpha\,\mathrm{var}(a) - \beta\,\mathrm{var}(b),$$

where $C \equiv E(b-a)$ represents the expected degree of conflict. To compare this with (3) and (4), I assume (without loss of generality) that $\alpha \geq \beta$, and expand the expression (3) for $EW(R^A)$:

$$(7)\quad EW(R^A) = -\beta\big(C^2 + \mathrm{var}(a) + \mathrm{var}(b) - 2\,\mathrm{cov}(a,b)\big).$$

Comparing (7) with (6), we see that *the rigid norm is the most efficient rule if and only if*

$$(8)\quad \beta^2 C^2 > (\alpha - \beta)\mathrm{var}(a) + 2\beta\,\mathrm{cov}(a,b).$$

From (8), we have, recalling that $\alpha + \beta = 1$:

PROPOSITION 2: *If $\alpha = \beta$, then the rigid norm is more efficient than the allocation of rights if and only if $C^2 > 4\,\mathrm{cov}(a,b)$.*

In particular, if a and b are uncorrelated, then the norm is the best rule of the three. The norm achieves a compromise, at the expense of responding to shifts in private information. So it is not surprising that its appeal is greatest when there is most conflict. When conflict is predominant, and there are increasing marginal costs to errors in the outcome (deviations from x^*), then it is more important to balance the average claims of conflicting interests than to respond flexibly to particular circumstances.

However, if the agents' preferences move together (positive covariance), then it is better (it may even be better for the nonchooser: see Bengt Holmström, 1984) to delegate the choice than to fix it in advance. Moreover, if α is much larger than β, then (provided $\mathrm{var}(a) > 0$, i.e., A's preferences are not fully known to the benevolent government) R^A outperforms R^n as well as R^B: *rights are efficient* (absent side payments) *if the outcome primarily concerns only one agent, who has unpredictable preferences.*

Without side payments, rights can enhance efficiency (relative to central benevolent decisions) when (*i*) the decision matters primarily to one person, and his (her) preferences are private information; or (*ii*) the central authority is ignorant of important facts that affect all concerned parties' preferences in the same direction, producing covariance in a and b. But in a symmetric problem in which the parties' preferences are independent, benevolent central decisions are more efficient than the result of allowing one interested party a right to choose.

III. Conclusion

I have compared the efficiency of some simple and commonly used rules for resolving conflicts. In a simple model, I showed that if one party to a conflict is much more concerned with the outcome than is the other, and his (her) preferences are unpredictable to a central authority, then he or she should have the right to determine the outcome. In a symmetric conflict, however, where two opposed parties care equally about a decision, I showed that unless their preferences are highly correlated, it is better to have a benevolent and disinterested central authority dictate the outcome, even if she is ignorant of the parties' precise preferences.

Claude d'Aspremont and Louis Gérard-Varet (1979) showed how, in such problems of social choice under incomplete information, a judiciously arranged scheme of announcements and side payments can implement the "first-best" outcome (here, x^*). Ronald Coase (1960) discussed how the simple allocation of rights, and side payments, can solve "externality" problems such as mine; however, the result does not carry over to the case where the agents themselves are imperfectly informed about one another's payoffs (see my earlier paper, 1986). Holmström has also discussed similar schemes in the context of principal-agent relationships.

While my model is clearly very special, and it would be foolish to draw policy conclusions directly from it, I believe that the basic tradeoff between desirable flexibility and desirable disinterestedness is a fundamental one, which will remain present in more sophisticated models as they are developed.

REFERENCES

Coase, Ronald, "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

Dasgupta, Partha, "Decentralization and Rights," *Economica*, May 1980, *47*, 107–23.

d'Aspremont, Claude and Gérard-Varet, Louis-Andre, "Incentives and Incomplete Information," *Journal of Public Economics*, February 1979, *11*, 25–45.

Farrell, Joseph, "Rights and Efficiency," mimeo., GTE Laboratories, 1986.

Holmström, Bengt, "On the Theory of Delegation," in M. Boyer and R. Kihlstrom, eds., *Bayesian Models in Economic Theory*, Amsterdam: Elsevier North-Holland, 1984.

Searching for Leviathan: Comment and Extension

By MICHAEL A. NELSON*

In a recent article, Wallace Oates (1985) investigated whether fiscal decentralization tended to act as a constraining influence on the overall size of the public sector. This hypothesis has been advanced by several alternative theories of government behavior including Richard Musgrave's (1959) model of how the "distribution" function of government would be carried out by subnational governments, Geoffrey Brennan and James Buchanan's (1980) "Leviathan" model, and the more traditional public choice model (for example, Walter Hettich and Stanley Winer, 1984). All contend that the decentralization of tax and spending decisions introduces competition among governmental units seeking to attract citizens and other mobile resources, and thereby constrains taxing power.

Oates analyzed two data sets, one pertaining to the state and local sector in the United States and the other consisting of a sample of 43 countries, and found little empirical support for the decentralization hypothesis. The purpose of this paper is to reconsider his analysis in the case of state and local governments. Specifically, it will be argued that Oates's measures of decentralization may be inappropriate to test the hypothesis. Furthermore, by taking into account the structure or composition of local governmental jurisdictions within a state, empirical support of the decentralization hypothesis can be marshaled for general-purpose local governments such as counties and municipalities.

I. Measuring Fiscal Decentralization Among the States

Centralization of tax and spending decisions has at least two dimensions in the state and local sector. One dimension relates to the division of functional responsibilities between the state and the local level; the other dimension relates to the degree to which decisions regarding locally assigned functions are decentralized locally. In his analysis, Oates alternatively employs two indices which are intended to measure fiscal decentralization in accordance with the first dimension: the state share of state-local general expenditures and the state share of state-local general revenues. However, data on each state's share of total expenditures or revenues may not accurately reflect the variation among states with respect to the division of functional responsibilities between the state and local level.

The problem is that preferences for the various types of public services provided by subnational governments, as well as the level of government assigned to provide each service, can influence the value of a state's expenditure and revenue shares. For example, consider states which have relatively strong preferences for elementary and secondary education, a service traditionally provided by localities. Other things being equal, these states are likely to have relatively lower state expenditure (revenue) shares due to higher spending (revenues generated) by localities for education. For these states, greater decentralization is implied relative to other states (which have less strong preferences for local schools), even when the division of responsibilities between the state and local levels is the same for both groups of states.[1]

*Department of Economics, Illinois State University, Normal, IL 61761. This research was done while on leave at Marquette University. I am grateful to Wallace Oates for many helpful comments and for providing an unpublished manuscript. I have also benefited from the comments of William Hunter, Peggy Nelson, and an anonymous referee.

[1] Even if preferences are homogeneous across states, expenditure and revenue data will not always capture

Oates's third index of fiscal decentralization is the absolute number of local governments within a state. This variable is presumably intended to measure the degree of decentralization regarding the discharge of local functions. This approach assumes that each type of local government within a state has a similar influence on public sector size. This assumption may not be realistic given that the scope of a jurisdiction's powers and operations varies substantially at the local level. For example, the U.S. Bureau of Census (July 1978, pp. 4–6) reports that nearly one-third of all local governments in the United States are special districts, although the proportion varies considerably from state to state. Most special districts are established to perform a single function; nearly one-half do not have authority to levy taxes and two-thirds have no full-time employees. Many townships also perform a very limited range of services.

Jurisdictions with limited functions (hereinafter referred to as single-function jurisdictions) may not be directly comparable with more general-purpose governments (for example, counties, municipalities) in an investigation of the decentralization hypothesis for two reasons. First, the hypothesis is based on the mobility of residents among competing governmental units. The constraining power of this mobility is likely to be greatly reduced in the case of single-function jurisdictions, especially if the service being provided is relatively minor (for example, cemetery special districts). The decentralization hypothesis seem most relevant for general-purpose local governments where the benefits from moving may be more substantial.[2]

Second, even if single-function jurisdictions are forced to compete for residents, it has been argued that general-purpose jurisdictions can provide local public services in a more cost-efficient manner. For example, the centralized administration of general-purpose governments may avoid the potential duplication of management activities. In addition, single-function governmental units may be too small to exhaust scale economies.[3] If the fiscal restraint potentially associated with more competition is dominated by a higher-cost service provision, then a positive relationship between public sector size and the number of single-function governments within a state might be expected.

In the following sections, how significant this dichotomy of local governments is with regard to the decentralization hypothesis will be investigated empirically.

II. The Decentralization Hypothesis Reconsidered

Given the inherent difficulty in measuring the division of functional responsibilities between state and local levels, the present analysis concentrates on how decentralized decision making at the local level affects public sector size. Specifically, does the type (i.e., single-function or general-purpose) as well as the number of local governments in a state make a difference in the size of the public sector in a state? Two measures relating to the size of the state-local sector are specifically analyzed:

G = State and local taxes as a fraction of personal income. This is the measure used by Oates.

G^* = State and local "adjusted" expenditures as a fraction of personal income.

whether state or local governments actually have the responsibility to make taxing and spending decisions for certain major functional responsibilities. In the case of public assistance, for example, Aid to Families with Dependent Children (AFDC) eligibility requirements and benefit levels are set by the state and are uniform statewide. Yet, the administration of the program, as well as the financing of the nonfederal share, varies from state to state. Oates's expenditure (revenue) share measure would indicate greater decentralization for those states where AFDC is administered (financed) locally. Similar problems exist in the areas of health and hospitals (for example, Medicaid).

[2] I am indebted to an anonymous referee for making this point.

[3] For a further discussion of these points, see the Advisory Commission on Intergovernmental Relations (ACIR, 1964). The view that general-purpose local governments offer cost advantages over single-function jurisdictions is not accepted by some public choice scholars. See Kevin Deno and Stephen Mehay (1985) for a summary.

No attempt is made with the first measure to account for variations among the states in the division of responsibilities between state and local governments and the influence that division could have on state-local sector size. With the second measure, these variations are controlled for by considering expenditures on only those functions that are uniformly the responsibility of either the state or the local level of government.[4]

Since this analysis concentrates on decentralization at the local level, a third measure of public sector size that relates strictly to localities is also considered. Because the functional responsibilities of local governments vary across states, a public service provided by the same general-purpose, local government nationwide is selected for analysis. The only major local function that satisfies these requirements for most states is fire protection. According to the ACIR (1982), the municipality is the dominant provider of fire services in all but two states. Accordingly, the third "size" measure considered is

Fire = Local expenditures on fire protection as a fraction of personal income.

An examination of this variable may offer insight into the decentralization issue that the other two size measures cannot. That is, if the specific type of government that provides some public service is held constant, will increasing the number of this type of jurisdiction (for a given population) have any effect on the level of expenditures?

While the role of decentralization between the state and local level with regard to which has the authority to discharge subnational responsibilities is not considered in this analysis, another related area of "centralized" decision making at the state level is. This pertains to mandated expenditures that a state places on its local governments. These mandates require that the local government either undertake a certain activity or meet a minimum standard. Even responsibilities carried out entirely by local governments can be affected by these mandates. Expenditures on local education, for example, can be affected by state-mandated special education programs or collective bargaining rules. The hypothesis is that these mandates may lead to a net increase in expenditures by local governments, and hence, the overall size of the public sector.[5]

In the next section, the relationship between public sector size, the relative use of state expenditure mandates, and local government structure, will be examined.

III. The Results

The analysis in this section is similar to that undertaken by Oates. To facilitate comparisons with Oates's results, the tax and expenditure data pertain to fiscal year 1977. In Table 1, simple Pearson correlation coefficients are reported for the two state-local size measures (G and G^*) and the number of local governments within a state by type. *General-Purpose* local governments are defined to include counties, municipalities, and where appropriate, townships.[6] *Single-Function* local governments include all other types of local governments including special districts, school districts and townships not classified as general-purpose jurisdictions. The number of local governments is normalized by population, for example, *General-Purpose* refers to average population per general-purpose government.

The results show that when local governments are considered in aggregate, states with higher average population per local govern-

[4] The major functions included in G^* are local schools, public welfare, fire protection, sanitation, and local parks.

[5] State restrictions placed on local governments with regard to the use of certain tax bases, tax rate limitations, and access to debt finance are not considered. The evidence from previous research on state-imposed restraints on local governments (ACIR, 1977, p. 3) suggests that, at least for the property tax, these restraints have no effect on the level of state and local spending.

[6] The functional responsibilities of townships closely resemble municipalities in New Jersey, Pennsylvania, and in the New England states (Bureau of the Census, July 1978, p. 3). In the remaining states, townships are classified as single-function governments.

TABLE 1—FISCAL DECENTRALIZATION AND STATE-LOCAL SECTOR SIZE: PEARSON CORRELATION COEFFICIENTS[a]

Decentralization Measure	State and Local Sector Size Variable		
	G	*G**	*Fire*[d]
All Local Governments[b]	.03	.09	–
General-Purpose[c]	.45[e]	.29[f]	.44[e]
Single-Function	–.08	.05	–
State Mandates to Local Governments	.51[e]	.38[f]	.42[e]

[a] G = Taxes; G^* = Adjusted Expenditures; *Fire* = Fire Protection Expenditures.

[b] Absolute number of local government units of this type within the state normalized by population. The states of Alaska and Hawaii are excluded from the calculations.

[c] For the correlation with *Fire*, general-purpose governments are restricted to municipalities and selected townships.

[d] Since general-purpose governments (municipalities) are the dominant provider of this service, correlations with the other government structure variables are not meaningful for this size measure. The states of Maryland and Vermont are excluded because the dominant service provider of fire protection in these states is not the municipality.

[e] Significance levels (two-tailed test) are 99 percent.

[f] Significance levels (two-tailed test) are 95 percent.

ment unit tended to have a larger state-local sector. While this finding is consistent with the decentralization hypothesis, the coefficients are not statistically different from zero in either case. The hypothesis that more local governments have no effect on public sector size cannot be rejected.

The relationship between public sector size and the type of local government jurisdiction is considered next. A positive, statistically significant, correlation is observed between the *General-Purpose* variable and both state-local size variables. States with a greater number of general-purpose local governments, relative to their population, tend to have a smaller state and local sector. For these governments, at least, the evidence supports the decentralization hypothesis. A similar result occurs when the analysis is restricted to an examination of fire protection expenditures at the local level. That is, *Fire* is positively associated with fewer municipalities, the dominant provider of this service.[7]

Turning to the single-function local government units, there is no evidence that the number of these jurisdictions has any effect on the size of the public sector. The correlation coefficient between the *Single-Function* variable and both G and G^* is statistically insignificant. These results suggest either that the decentralization hypothesis is not relevant for these types of governments or the higher-cost service provision of these units outweighs any revenue-constraining effects of greater decentralization.

The last line of Table 1 shows the correlation between the absolute number of state expenditure mandates on local governments and the public sector size variables.[8] In all three cases, a positive and statistically significant coefficient is found. Greater centralization of decision making via state mandating tends to be directly related to the size of the state and local sector.

Following Oates, multiple regression analysis is performed next to control for other factors that could affect the size of the public sector. Three equations are estimated for each dependent variable and the results are reported in Table 2. In equation (1), each size variable is regressed against the normalized local government structure variables.[9] For all three size measures, the positive sign of the *General-Purpose* government variable

[7] An anonymous referee points out that the positive relationship between fire protection expenditures and population per jurisdiction may simply reflect diseconomies of scale. Studies of scale economies in the case of fire protection vary as to whether average unit cost declines continually with population or is U shaped. See Werner Hirsch (1970, pp. 182–84).

[8] The ACIR (1978) gathered information on the existence of mandates in 77 functional areas covering seven categories of local expenditures by surveying state and local officials. If "no response" was reported by the ACIR in more than one of the seven categories, that state was deleted from the present analysis. Generally complete data are available for only 35 states in the case of the state-local size variables and 39 states for local fire protection.

[9] Following Oates, a logistic transformation is made on all dependent variables prior to estimation. The results are essentially the same as those when the transformation is not made, a finding similar to Oates.

TABLE 2—THE EFFECT OF FISCAL DECENTRALIZATION ON THE SIZE OF THE STATE-LOCAL SECTOR: RESULTS OF REGRESSIONS[a]

Indepent Variable	Dependent Variable								
	G			*G**			*Fire*[d]		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Intercept	−2.05	−2.07	−2.43	−1.84	−2.03	−2.60	−6.10	−6.67	−6.54
	(67.91)	(5.81)	(4.88)	(56.35)	(5.67)	(5.31)	(85.68)	(17.99)	(17.70)
Local Government Structure[c]									
General-Purpose	.008[e]	.012[e]	.009[f]	.005[g]	.009[f]	.004	.016[e]	.010[g]	.014[f]
	(3.50)	(3.37)	(2.25)	(1.93)	(2.44)	(0.91)	(3.28)	(1.82)	(2.28)
Single-Function	−.001	−.0006	−.0005	−.00001	.00007	.002[g]			
	(1.39)	(1.07)	(0.63)	(0.03)	(0.12)	(1.80)			
Other Explanatory Variables[d]									
State Mandates			.008[f]			.010[f]			.057[e]
			(2.34)			(2.71)			(3.01)
Per Capita Personal Income		.00004	.00004		.00002	.00003		.00004	−.00004
		(1.17)	(0.89)		(0.52)	(0.64)		(0.61)	(0.66)
Intergovernmental Grants		−0.42	−0.12		0.47	0.90			
		(0.56)	(0.12)		(0.86)	(1.29)			
Population		−.000002	−.000005		−.0000005	−.0000005		−.00002	−.00003
		(0.36)	(0.70)		(0.07)	(0.06)		(1.67)	(1.80)
Urban[b]		−.003[f]	−.003[g]		−.002	−.002		0.008[e]	0.008[e]
		(2.24)	(1.77)		(1.47)	(1.14)		(3.39)	(3.27)
R-Squared	.22	.33	.45	.08	.19	.35	.20	.41	.52
F-Statistic	6.32	3.44	3.17	1.96	1.60	2.11	10.77	7.04	7.04
S.E.E.	.13	.13	.13	.14	.14	.14	.32	.28	.25
Number of Observations	48	48	35	48	48	35	46	46	39

[a, e, f]: See Table 1.

[b] Percent of population residing within SMSAs.

[c] For *Fire*, the dominant service provider is general-purpose governments (i.e., municipalities) for this function; single-function local governments are not relevant.

[d] For *Fire*, grants are omitted due to lack of data.

[g] Significance levels (two-tailed test) are 90 percent.

supports the decentralization hypothesis, although the evidence is statistically weak in the case of G^*. On the other hand, for single-function jurisdictions, the coefficient is not statistically different from zero in either of the state-local sector size equations where this variable is specified.

In equation (2), the four additional variables used by Oates to control for other influences on the size of the public sector (*Per Capita Personal Income*, *Intergovernmental Grants*, *Population*, and *Urban*, the percentage of population residing within SMSAs) are added to the local governmental structure explanatory variables. The overall explanatory power of this equation is considerably improved over the first equation. More importantly, in all cases there is statistically significant evidence in support of the decentralization hypothesis for general-purpose governments.[10]

Finally, in equation (3), the effects of state expenditure mandates on local governments

[10] In the case of the size measure G^*, the *Grants* variable was adjusted to exclude federal grants for those functions deleted from this size measure. The *Mandates* variable was similarly adjusted.

are considered. For all three dependent variables, the hypothesis that the greater use of expenditure mandates will lead to a larger public sector is strongly supported.

IV. Conclusions

Further evidence has been presented on the relationship between fiscal decentralization, as measured by the number of residents per jurisdiction, and public sector size. The current research is distinguished from Oates's work in terms of the specific dimension of decentralization that is analyzed. This study has focused on the effects of greater decentralization by increasing the number of local governments for a given population. For the most part, Oates investigated the effects of greater decentralization from shifting responsibilities to lower levels of government.

Fairly robust empirical evidence was presented in support of the decentralization hypothesis in the case of general-purpose, local governments. However, more substate jurisdictions for a given population will not always contribute to a smaller public sector. Specifically, increasing the number of jurisdictions which are established to perform only a very limited number of functions, such as most special districts, seem to have little effect on public sector size.

These findings are consistent with a basic tenent of the Leviathan model, which holds that greater decentralization results in increased competition among subnational governments. This competition acts to constrain a revenue-maximizing, monolithic government that operates in a less competitive environment. An analysis of the local sector is especially appropriate to test the Leviathan proposition, since competition among jurisdictions is most likely to be observed at the local or substate level, where mobility among jurisdictions is relatively easy.

The results of this paper, however, are just as consistent with the predicted outcomes of other models of government behavior, where competition among jurisdictions is assumed, but the public sector is not characterized as a revenue-maximizing monolith. For example, the findings are compatible with models which predict that decentralization of public welfare will result in a comparatively smaller budget for that function.

Finally, the findings are compatible with explanations of government behavior not related to competition among jurisdictions. In particular, Oates (1986) has argued that publicly provided goods with significant indivisibilities, where a large expenditure is necessary to provide the first unit of the good, cannot be efficiently provided unless the population is sufficiently large. Citing earlier work by Henry Schmandt and G. Ross Stephens (1960), he argues that bigger cities may provide a greater range of public goods, with more expenditures per resident, than less populated jurisdictions.[11]

In summary, this paper has provided relatively strong evidence that a centralized local sector is associated with a higher level of public spending relative to income. Whether this result is due to Leviathan forces operating in a monopolistic environment, or some other phenomena, is a question to be addressed by future research.

DATA APPENDIX

Data on state-local taxes, expenditures, intergovernmental grants and income were taken from: Bureau of the Census (November 1978, Tables 6, 13, and 27); population and the number of local governments: Bureau of the Census (July 1978, Tables 2 and 3); urban: U.S. Department of Commerce (1979, p. 19); and mandates: ACIR (1978, Table III-4).

[11] In a related paper, James Litvack and Oates (1970, pp. 51–52) have also argued that the price of impure public goods is likely to rise as population becomes highly concentrated due to higher congestion costs. Assuming that the demand for these goods is price inelastic, then per capita expenditures on these goods would be expected to increase with population density.

REFERENCES

Brennan, Geoffrey and Buchanan, James, *The Power to Tax: Analytical Foundations of a Fiscal Constitution*, Cambridge; New York: Cambridge University Press, 1980.

Deno, Kevin T. and Mehay, Stephen L., "Institutional Constraints on Local Jurisdiction Formation," *Public Finance Quarterly*, October 1985, *13*, 450–463.

Hettich, Walter and Winer, Stanley, "A Positive Model of Tax Structure," *Journal of Public Economics*, June 1984, *24*, 67–87.

Hirsch, Werner Z., *The Economics of State and Local Government*, New York: McGraw-Hill, 1970.

Litvack, James and Oates, Wallace, "Group Size and the Output of Public Goods: Theory and an Application to State-Local Finance in the United States," *Public Finance*, 1970, *25*, 42–62.

Musgrave, Richard A., *The Theory of Public Finance*, New York: McGraw-Hill, 1959.

Nelson, Michael, "An Empirical Analysis of State and Local Tax Structure in the Context of the Leviathan Model of Government," *Public Choice*, 1986, *49*, 283–94.

Oates, Wallace, "Searching for Leviathan," *American Economic Review*, September 1985, *75*, 748–57.

______, "On the Measurement of Congestion in the Provision of Local Public Goods," unpublished manuscript, January 1986.

Schmandt, Henry and Stephens, G. Ross, "Measuring Municipal Output," *National Tax Journal*, December 1960, *13*, 369–75.

Advisory Commission on Intergovernmental Relations, *The Problem of Special Districts in American Government*, Washington: USGPO, 1964.

______, *State Limitations on Local Taxes and Expenditures*, Washington: USGPO, 1977.

______, *State Mandating of Local Expenditures*, Washington: USGPO, 1978.

______, *State and Local Roles in the Federal System*, Washington: USGPO, 1982.

U.S. Bureau of the Census, *1977 Census of Governments*, Vol. 1, No. 1: *Governmental Organization*, Washington: USGPO, July 1978.

______, *Governmental Finances in 1976–77*, Washington: USGPO, November 1978.

______, *1977 Census of Governments*, Vol. 4, No. 2: *Finances of Special Districts*, Washington: USGPO, May 1979.

U.S. Department of Commerce, *Statistical Abstract*, Washington: USGPO, 1979.

U.S. Department of Health and Human Services, *Social Security Bulletin*, *Annual Statistical Supplement 1977–79*, Washington: USGPO, 1980.

The Validity of Studies with Line of Business Data: Comment

By F. M. Scherer, William F. Long, Stephen Martin, Dennis C. Mueller, George Pascoe, David J. Ravenscraft, John T. Scott, and Leonard W. Weiss*

In the March 1985 issue of this *Review*, George Benston found fault with Federal Trade Commission Line of Business (*LB*) data generally and singled out for extended criticism thirteen *LB* data-based papers written by the authors of this comment. Even by the pre-Queensberry rules governing economic disputation, Benston's article is one-sided and negative. Moreover, it is marred by numerous errors in characterizing our work. We wish to set the record straight.

I. The Line of Business Research Program

The data on which Benston focuses are financial performance statistics collected by the FTC for the years 1974–77 from some 437 to 471 U.S. corporations. The data are disaggregated to "lines of business" defined at a level of manufacturing industry detail ranging between the 3- and 4-digit divisions of the Standard Industrial Classification.

Benston's main theme is his doubt "whether meaningful empirical work on the relationships between [market] structure and performance is feasible" (p. 64). However, most of his critique could pertain to any systematic use of accounting data, and the studies he explicitly attacks sweep a much wider range of economic phenomena. The data have been used to analyze research and development (*R&D*)—productivity links, transfer pricing by multinational enterprises, business diversification strategies, the profitability of mergers and sell-offs, the relationship between profitability and stock market risk, how federal *R&D* contracts affect private *R&D* expenditures, and much else. As of February 1986, research with the disaggregated data had yielded some 63 papers, nearly half of which were published or accepted for publication in refereed journals and compendia.[1] In addition, an unknown but substantial number of papers have used the published *LB* industry aggregates.

Benston correctly observes that one motivation for the Line of Business program was the perceived inadequacy of existing data for ascertaining how market structure, and, in particular, seller concentration, affected profitability (taken as an index of industry performance). However, he fails to recognize the impact *LB* data analyses have had in modifying views on such structure-performance relationships. At the time the *LB* program was initiated, the state of knowledge was characterized in an exchange between Harold Demsetz (1974) and Weiss (1974). Weiss pointed to the large number of empirical studies demonstrating that profitability rose systematically with seller concentration, while Demsetz questioned whether the chain of causation reflected the ability of the larger firms in concentrated markets to maintain elevated prices, or their greater success in reducing unit costs or securing other advantages. The contending schools of thought were deadlocked; only with better data could the impasse be resolved. Subsequent simulations (Ravenscraft, 1984) confirmed Demsetz' intuition that when leading sellers have unit cost advantages over smaller firms, regression analyses at the industry level could, because

*Scherer: Swarthmore College; Long, Pascoe, and Ravenscraft, Federal Trade Commission Bureau of Economics; Martin, Michigan State University; Mueller, University of Maryland; Scott, Dartmouth College; Weiss, University of Wisconsin. All have been affiliated with the FTC Line of Business program as staff economists or consultants. A review by FTC staff has determined that individual company Line of Business data are not disclosed in this comment. The views here are our own and not necessarily those of the FTC.

[1] Some of the papers referred to by Benston, and others cited here, are not published. All that are listed as FTC working papers can be obtained upon request from the FTC Line of Business office.

of aggregation biases, yield positive profit-concentration coefficients even when no price-raising effect was present. To avoid the biases, it is necessary to analyze performance at the individual line of business level and to include market share as well as industry concentration variables.

The research with *LB* data has gone forward.[2] It has shown that individual market share effects are indeed much more powerful than the traditionally emphasized concentration effects in explaining profitability. With most specifications, concentration coefficients turn out to be *negative*, not positive, in conjunction with market share variables. The positive and significant market share relationships alone cannot discriminate between monopoly power and efficiency or cost advantage hypotheses. However, Ravenscraft (1983) has shown that the market share measures interact with capital intensity, implying scale economies, and advertising, which reflects product differentiation effects.

A more recent analysis of *LB* data by John Kwoka and Ravenscraft (1986) reveals that the profit advantage of larger sellers depends upon structural variables which in turn influence how the market leader chooses to exploit its strategic advantages. Analyses by Bradley Gale and Ben Branch (1982) of non-FTC data disaggregated to the line of business level show the market share-profit relationships to follow from a mixture of cost, first mover, and subjectively determined quality advantages enjoyed by larger sellers. Work by Scott (1982) and Scott and Pascoe (1986) has suggested that industrywide seller concentration operates in more subtle ways dependent inter alia upon relative bargaining position and capital intensity.

II. Accounting Biases?

Although Benston does not acknowledge how much economists' understanding of structure-profit relationships has changed, he is aware that strongly positive market share effects have been a prominent feature of the new analyses. He attempts to dismiss them (pp. 40–41 and 56–57) as the possible consequence of accounting biases or other anomalies correlated with market share. To the extent that evidence is cited, it is done selectively. Thus (at pp. 39–40) he cites empirical studies showing systematic relationships between the choice of profit-altering accounting methods and firm size, company control form, and the presence of debt covenants. He fails to note that one of his cited studies (Robert Hagerman and Mark Zmijewski, 1979) found that firms in more concentrated markets tended to choose accounting policies that *reduced* their reported profits[3]—the opposite of what one would expect if the positive coefficients estimated for concentration (earlier) and market share (more recently) stemmed from accounting bias. And to the extent that firm size is related to market share (which need not occur in a multi-industry company), the observed tendency for larger corporations to prefer profit-reducing accounting policies must weaken, not strengthen, the relationships obtained in structure-profitability studies.

Benston's main methodology, however, is worst-case analysis. After reciting a litany of well-known problems that can infiltrate accounting data, he assumes without evidence that if anything can go wrong, it will. He thus ignores a corollary to Murphy's Law that might be called Leamer's Lemma (1985): If something can go wrong, one has an obligation to perform sensitivity tests to see whether it actually has, and if so, what the range of effects is. Indeed, for exploring the impact of accounting convention choices on structural relationships, few vehicles are more suitable than the *LB* data base, which contains explicit information on the amount and method of common cost and asset allocations, transfer pricing methods and magnitudes, inventory accounting methods, depreciation conventions, depreciable asset ages, merger accounting methods,[4] and much else.

[2] To be sure, other recent research using more aggregated data has helped clarify the debate over structure-performance relationships. For a wider-ranging survey, see Ravenscraft (1984).

[3] See also Hagerman and Zmijewski (1981).

[4] The information on merger accounting methods was not in the original *LB* data base, but was linked to it from outside sources, contradicting Benston's asser-

A. *Profit-Structure Sensitivity Tests*

The richest set of sensitivity tests thus far has been carried out in the context of the profit-market structure models Benston criticizes. Table 1 provides a summary, focusing on the coefficients estimated for market share and (in a more limited set of cases) four-firm seller concentration ratios across 15 studies varying widely in profit variable definition, sample year, sample composition, model structure, and control for accounting method choices.[5] It excludes models like Martin's (1983), in which the market share variable is endogenous, but continues to be significantly positive, or those in which the structure variables interact with other variables, yielding quite different interpretations of profit function partial derivatives. It encompasses, however, a diversity of approaches to the extreme value problems encountered in using *LB* data—for example, when profit ratios in the minus 100 percent range occur because of new business startups, accidents, or impending exit.

The first four equations report the market share and concentration coefficient changes for different years using Ravenscraft's (1983) basic operating income/sales regression with the most inclusive *LB* sample manufacturing line coverage. Earlier research at the industry level showed seller concentration coefficients to vary with the business cycle, being most strongly positive in the late 1950's and early 1960's and weakest in the inflationary 1970's. See Weiss (pp. 200–03, 221). Mild business cycle effects are evident here too. The smallest market share effect is for 1974, when price controls gave way to soaring inflation and then the start of a recession. The largest effect is for recovery year 1977. In every year, the market share coefficients show that profitability doubled or even trebled with increases in market share over the range of observed values. Four-firm seller concentration coefficients remain negative, small in comparison to analogously scaled market share effects, and hovering near statistical insignificance.

Equations (5) and (6) compare essentially similar regressions, except that in equation (6), Ravenscraft (1981) adjusted all inter-*LB* transfers made at nonmarket prices to a basis consistent with those determined for *LB*s in the same industry transferring at market prices. Glejser heteroskedasticity corrections were also implemented. The market share and concentration coefficients do not change appreciably. Benston acknowledged this result (p. 49), but forgets it in arguing that the strongly positive market share coefficients obtained by Ravenscraft in an analogue of Table 1's equation (2) might have resulted because "possibly transfers to those units were made at less than market prices" (p. 57).

Equation (7), which has not been reported previously, shows how the Ravenscraft results for 1975 (equation (2)) change when common or "nontraceable" costs are reallocated from the corporate pool on the basis of "market" rates—that is, reflecting the cost burdens of otherwise comparable lines with *no* nontraceable cost allocations. The market share coefficient drops by a third but remains strong and significant; the con-

tion "that firm-specific information from other sources cannot be brought in" (p. 51). Many other external data linkages—for example, involving patents received, input-output statistics, stock prices, merger activity, company internal organization and management turnover, etc.—have been accomplished.

[5] The implications of our Table 1 are strikingly different from those of Benston's Table 3 (p. 57), in which he reports estimates for some coefficients, including those for market structure. Most of the large differences shown in his table (whose year headings are erroneously reversed) stem from Benston's failure correctly to adjust variable scalings—for example, when market shares or concentration indices were measured as ratios by one author and as percentages by another. The sign on the Weiss-Pascoe market share coefficient is incorrectly reported; when this error is corrected, all signs are identical. Many of the variables in the table are not in fact identically defined, contrary to the table's implication. For example, Ravenscraft's asset/sales variables were adjusted for capacity utilization, whereas Martin's were not. Other coefficients differ because of the inclusion or exclusion of additional terms interacting with them (which alters their economic interpretation) and collinear variables (such as a cost disadvantage ratio, coupled with the minimum efficient scale variable by Martin, but not by Ravenscraft). Benston incorrectly identifies the source of Martin's estimates; the paper he cites presents a five-equation simultaneous model; the coefficients Benston reproduces are from an earlier four-equation model.

TABLE 1—EFFECTS OF YEAR DIFFERENCES AND ACCOUNTING METHODS ON STRUCTURE-PERFORMANCE RESULTS

Equations	Special Emphasis	No. of Obs.[a]	Year	Dependent Variable[b]	Market Share	Concentration	Other Variables	R^2
(1) Ravenscraft (1983)[c]		3030[d]	1974	*OI/S* [.073]	.1793 (4.71) [.038]	−.0370 (−2.22) [.387]	19 other *LB*, firm and industry variables	.223
(2) Ravenscraft (1983)		3186	1975	*OI/S* [.065]	.1833 (4.90) [.037]	−.0218 (−1.34) [.387]	21 other *LB*, firm and industry variables	.208
(3) Ravenscraft (1983)[c]		3185	1976	*OI/S* [.071]	.2162 (5.13) [.036]	−.0350 (−1.94) [.387]	21 other *LB*, firm and industry variables	.160
(4) New regression for this comment		2955[e]	1977	*OI/S* [.078]	.2335 (7.60) [.037]	−.0343 (−2.60) [.391]	15 other *LB*, firm and industry variables	.067
(5) Ravenscraft (1983)	Heteroskedasticity correction	3186	1975	*OI/S* [.065]	.1476 (5.51) [.037]	−.0222 (−1.77) [.387]	21 other *LB*, firm and industry variables	.128
(6) Ravenscraft (1981)	All transfers at market prices, heteroskedasticity correction	3004[d]	1975	*OI/S* [.066]	.1432 (5.33) [.038]	−.0195 (−1.54) [.388]	29 other *LB*, firm, and ind. variables, including transfer method controls	.154
(7) New regression for this comment	All common costs allocated at market rates	3186	1975	*OI/S* [.068]	.1234 (3.18) [.037]	−.0240 (−1.40) [.387]	21 other *LB*, firm and industry variables	.150
(8) Long and Ravenscraft (1984)[c]	Depreciation charges not subtracted from numerator	3014[d]	1975	*OI + Dep/S* [.092]	.2000 (5.39) [.038]	−.0220 (−1.34) [.390]	21 other *LB*, firm and industry variables	.168
(9) Benvignati (1986)	Estimated capital costs subtracted from numerator	2635[f]	1975	*OI-KCst/S* [.018]	.2620 (5.95) [.037]	–[g]	16 other *LB*, firm and industry variables	.282
(10) Long and Ravenscraft (1984)[c]	Assets instead of sales used in denominator	3014[d]	1975	*OI/A* [.092]	.1720 (2.73) [.038]	.0015 (0.05) [.390]	21 other *LB*, firm and industry variables	.134
(11) Schmalensee (1985)	Relative effects of firm, industry, and share	1775[h]	1975	*OI/A* [.137]	.2359 –[i] [.061]	–[j]	455 firm effect and 241 industry industry effect dummy variables	.496
(12) Ravenscraft and Scherer (1986a)	Merger analysis	2955[e]	1977	*OI/A* [.139]	.3925 (6.34) [.037]	–[j]	257 industry effect dummy and 4 merger variables	.155
(13) Ravenscraft and Scherer (1986a)	Merger analysis, effects of accounting methods	2955[e]	1977	*OI/A* [.139]	.3691 (5.91) [.037]	–[j]	257 industry effect dummy and 4 merger and 2 accounting variables	.159
(14) Marshall (1986)	Comparison of PIMS and *LB*: *LB* data used	2450[k]	1974–77	*TrOI/TrA* [.190]	.2709 (5.14) [.041]	.0127 (0.63) [.390]	7 other *LB* and industry variables	.131
(15) Marshall (1986)	Comparison of PIMS and *LB*: PIMS data used	837[k]	1974–77	*TrOI/TrA* [.192]	.3249 (9.82) [.246]	−.0185 (−0.73) [.559]	7 other *LB* and industry variables	.263

Sources: For all equations except (4) and (7), see the References. For equation (4) see Ravenscraft and Scherer (1986a). For equation (7), see Ravenscraft (1983) and Long et. al. (1982).

Notes: Means are in square brackets [ ], and t-values are shown in parentheses ( ).

[a] For an explanation for minor differences in number of observations among studies, see individual studies. For major differences, see fnn. d, e, f, and h below.

[b] Variables used to describe profitability measures: *OI*-Operating Income; *TrOI-OI*, traceable, i.e., only traceable operating costs are subtracted; *OI + Dep-OI* + depreciation charges; *OI-KCst-OI* less estimated capital costs; *S*-Sales; *A*-Assets; *TrA*-Assets, traceable.

[c] Discussed, but not presented.

[d] Excludes *LB*s from companies added in 1975.

[e] Excludes *LB*s with inadequate merger data; outliers.

[f] Excludes *LB*s with missing trade data.

[g] Not comparable, since interaction terms with concentration are used.

[h] Excludes *LB*s in industries designated as "miscellaneous" or "not elsewhere classified"; *LB*s with market share less than 1 percent.

[i] Significant at the 5 percent level.

[j] Cannot be included because of 4-digit industry effect dummies.

[k] Excludes *LB*s not in all four years; *LB*s with zero traceable assets; outliers.

centration coefficient is hardly affected. Equation (8) estimates the basic equation (2) model with depreciation added back into operating income. The effects on market share and concentration coefficients are small.

An even more stringent profit variable redefinition is covered by equation (9). Anita Benvignati (1987) subtracted from the operating income/sales ratios estimates by Scott and Pascoe (1984) of securities market-based capital costs, in effect specifying the dependent variable as a measure of excess returns. The market share coefficient remains robust.

Regressions (10)–(12) relate the ratio of operating income to assets, rather than sales, with a diverse assortment of additional controls. The market share coefficients are larger than in the sales regressions because mean operating income/assets ratios (reported in brackets) are larger than the comparable sales-deflated ratios. They are also more variable over the business cycle, although they remain powerful at all stages. Regression (10) is most similar to the earlier equations; there, for the first time, the seller concentration coefficient turns positive but insignificant.

Richard Schmalensee's (1985) regression (11) is the most different, culling out from the sample all lines in "miscellaneous" manufacturing categories and those with market shares of less than 1 percent, and controlling by means of fixed effect dummy variables for both company effects (with respect to which accounting choices are believed to differ) and industry effects. The market share coefficient remains in the same range as in quite differently specified models.[6]

Regressions (12) and (13) control identically for industry and merger accounting effects on 1977 profitability. They differ in the addition to equation (13) of variables measuring the extent to which LIFO inventory accounting and straightline (as contrasted to accelerated) depreciation methods were adopted. The straightline variable was significantly related to profitability, but the market share coefficient changes by only 6 percent with the added accounting method controls.

Equations (14) and (15) introduce what is without doubt the most glaring oversight in Benston's critique. The FTC's Line of Business program, on which his analysis focuses exclusively, has a private sector counterpart —the PIMS (Profit Impact of Market Strategies) program. Financed by several hundred participating companies, it too collects profit-and-loss and other accounting data at the line of business (i.e., "business unit") level. Thus, the kind of data collection Benston attacks as meaningless has passed a clear market test. Early analyses using PIMS data, notably, by Gale and Branch, yielded market share and seller concentration coefficient effects quite similar to those reported in our Table 1.

Equations (14) and (15) come from a study by Cheri Marshall (1986) estimating profit-structure relationships for Line of Business *and* PIMS data sets matched as exactly as possible with respect to business cycle phase, coverage of manufacturing lines only, and model specification. Even though the PIMS data tend to oversample leading sellers with high market shares and to define markets more narrowly than the *LB* program does, the market share coefficients estimated with these two data sets turn out to be quite similar. The concentration coefficients have different signs, but fall far short of statistical significance.

To sum up, contrary to the implications drawn by Benston, the basic structural relationships estimated using *LB* (and PIMS) data turn out to be robust across a wide range of variable definitions, sampling frames, and controls for accounting method variations.

B. *Other Accounting Impact Analyses*

Two further tests of how accounting choices matter were reviewed by Benston. He draws negative implications mainly by

[6] No industry concentration coefficient could be estimated, since it would be perfectly collinear with the fixed industry effect dummies.

quoting the authors out of context or by ignoring results that run contrary to his conclusions. Thus, summarizing Long's discussion (1982) of how the coefficients of a structure-performance model were changed by applying alternative cost allocation formulae, he emphasizes Long's observation that the effect of "random" allocation was "disasterous" (p. 48).[7] He fails to note the concluding sentence of the same paragraph, in which Long observes that the random allocation method had "serious problems." Contrary to Benston's assertion that it "has about as much validity as other methods," it places on average the same absolute burden on lines with 2 percent of a firm's sales as on those with 40 percent—a situation that is neither realistic nor competitively sustainable.[8] Likewise, despite citing Long's follow-up analysis (Long et al., 1982) in his references, he fails to acknowledge the insights from market-oriented cost allocation tests conducted by Long. Those tests showed the random allocation method to be by far the *least* consistent of several methods with a market-based procedure.

Benston correctly summarizes most of Ravenscraft's research (1981) on the effects of alternative transfer pricing methods. However, he singles out for quotation a remark by Ravenscraft indicating that the shift from nonmarket to market transfer methods changed operating income-sales ratios by as much as 17,595 percent, with an average of 445 percent. He fails to mention Ravenscraft's footnote, which observes that since most of the nonmarket transfers were at cost, the profitability denominator with respect to which these high percentage changes are calculated was typically close to zero. And of course, any finite change, however modest, divided by zero or something close to zero leads to very large numbers of the sort Ravenscraft reported.[9]

An accounting choice variable not explicitly considered by Benston is the determination of what asset values will be recorded after a merger occurs. There are two main alternatives: purchase accounting, under which acquired assets are revalued to reflect the actual transaction price paid, and pooling of interests, under which premerger book values are retained. Studies by Ravenscraft and Scherer (1986a) show that postmerger profit rates of acquisition-prone lines of business are strongly affected by this choice. Lines with substantial amounts of purchase accounting assets report significantly lower profitability postmerger than lines using pooling of interests accounting. However, this bias has been identified and broken down into components associated with the write-up of asset values, increased depreciation (typically modest), and a selection bias reflecting the tendency for purchase mergers to have poorer premerger profitability prospects. Thus, with the appropriate analytic effort, differences in accounting method choices can be turned into a lever for understanding better the economics of merger.

In sum, accounting method choices do make a difference in reported profit figures. However, it is possible to go well beyond Benston's speculation on what *might* happen, analyzing both the magnitude of their effects on reported profits and any systematic biases they might impart to structural coefficient estimates. The work done thus far refutes Benston's claim that the biases in structure-performance and other analyses are so serious as to vitiate the results. After elaborate and diverse controls, the basic relationships persist. As always, this cannot be the last word; more remains to be done.

[7] The misspelling was not in Long's original. It was enclosed in quotation marks in Benston's original (1982, p. 31) and hence survived subsequent editing.

[8] Even with the large changes in some profit-sales ratios induced by the random allocation procedure, structure-performance hypothesis tests were little affected. Of 8 coefficients that were positive and significant at the 10 percent level, 7 remained positive and significant after the reallocations. Of 6 that were negative and significant, 4 remained negative and significant after the change.

[9] The same "divide by zero" problem strongly influences the alternative common cost allocation results reported by Robert Mautz and Fred Skousen (1968), and cited approvingly by Benston (p. 48).

III. Too Much Variability?

Benston criticizes the Line of Business data among other reasons because reported profit figures exhibit too much variability. Observing that 16 percent of the lines in Martin's sample had negative profitability in (recession) year 1975, he suggests that the data reflect "very limited entry and exit into a substantial number of markets" or "substantial annual random variation in returns, which implies that the data measurement period should be longer than a year," or "the effects of accounting practices" (p. 53). Without intervening clarification, he later expresses his summary view that "these data reflect the accounting biases present in the numbers" and that "it is doubtful whether analyses using these data would yield valid findings" (p. 64). Similarly, Benston expresses concern over large year-to-year changes in profitability ratios for data aggregated to the industry level, suggesting that "...changes in the environment in which the reporting companies operated or changes in the sample caused these differences. The errors of measurement described above could also be responsible for the differences between years" (p. 52).

A. *A Prediction Test*

Recognizing that the measures of "true" economic value against which *LB* data might ideally be judged are not available, Benston argues that "The validity of company accounting data for economic analyses... would be best determined with tests of prediction" (p. 39).[10] He attempts no such test. However, a strong test is possible.

A fundamental proposition of economic theory is that firms exit a market in response to negative profits—that is, losses. Benston finds implausible the high incidence of losses in individual lines, but makes no attempt to test whether there is in fact "very limited entry and exit."

The characteristic form of exit in large manufacturing corporations is sell-off; few units are simply shut down. Among the roughly 4,000 manufacturing *LB*s operated by companies in the FTC sample, 436 lines were sold off totally, and at least 455 more experienced partial sell-off, between 1974 and 1981. For lines with no recorded sell-off activity, operating income (i.e., profit before deduction of interest charges, income taxes, and extraordinary items) averaged 13.93 percent of assets. For the lines that were totally sold off, the trend in operating income-assets ratios (in percentage terms) over the six years before sell-off (in year *T*) is shown in Table 2. The pattern is what one would expect if *LB* profitability figures were valid decision-making indicia. About three years before sell-off, profitability declines markedly (and statistically significantly), dropping further two years before sell-off and turning *negative* on average in the year before sell-off.

TABLE 2—PREDIVESTITURE PROFITABILITY TRENDS

Year[a]	Operating Income / Assets[b]
$T-6$	9.3
$T-5$	8.3
$T-4$	7.1
$T-3$	3.5
$T-2$	2.9
$T-1$	-0.3

Source: Ravenscraft and Scherer (1986b).
[a] *T* here is the year in which sell-off was initiated.
[b] Shown in percent.

The *LB* profitability variable was the most powerful single predictor in a logit regression equation whose dependent variable equals unity when a business unit was fully divested during the 1976–81 period.[11] Also, holding *LB* profitability constant, sell-off was significantly more likely, the lower overall *company* profitability was. It seems clear

[10] See also Milton Friedman (1953).

[11] See Ravenscraft and Scherer (1987, ch. 6.). Other statistically significant variables included past merger history, market share, *R&D* intensity, and whether there was a change in the chief executive officer in the two years preceding sell-off. An advertising intensity variable was not significant, presumably because advertising, unlike *R&D* and contrary to Benston's statement (p. 43), is known to depreciate rapidly in most instances. Compare Darral Clarke (1976).

TABLE 3—RANK AND PROFITABILITY CHANGES OF INDUSTRIES AMONG THE TEN MOST PROFITABLE IN AGGREGATED *LB* REPORTS FOR 1974, 1975, AND 1976

FTC Industry Code	Industry Description	Operating Income/Assets Rank (and Percentage)[b]		
		1974	1975	1976
33.04[a]	Primary lead	1 (41.8)	76 (15.0)	30 (21.3)
20.17[a]	Beet sugar	2 (41.5)	15 (25.5)	210 (5.8)
28.14[a]	Fertilizers	3 (40.3)	4 (36.4)	151 (11.6)
20.09	Cereal breakfast foods	4 (39.8)	3 (38.1)	3 (38.3)
28.15[a]	Pesticides and agricultural chemicals	5 (34.1)	7 (29.5)	49 (19.0)
28.08	Proprietary drugs	6 (31.0	30 (22.2)	9 (23.3)
29.03[a]	Misc. petroleum and coal products	7 (29.6)	39 (20.2)	192 (8.5)
26.01[a]	Pulp mills	8 (29.1)	47 (19.0)	66 (17.5)
27.06[a]	Manifold business forms	9 (28.9)	17 (24.1)	58 (18.1)
33.05[a]	Primary zinc	10 (27.3)	191 (5.8)	205 (6.2)
34.09[a]	Fabricated structural metal	223 (0.0)	1 (39.2)	1 (41.9)
34.03	Cutlery	20 (24.5)	2 (38.2)	2 (38.5)
20.13[a]	Wet corn milling	16 (25.0)	5 (35.0)	152 (11.5)
20.02[a]	Poultry and egg processing	232 (−11.6)	6 (29.5)	226 (1.1)
26.08	Stationery, tablets, etc.	24 (22.5)	8 (28.9)	42 (19.8)
20.15	Cookies and crackers	41 (18.7)	9 (28.7)	4 (33.9)
35.26	Speed changers and industrial drives	46 (18.2)	10 (28.0)	8 (27.9)
21.03	Chewing and smoking tobacco	11 (27.2)	13 (26.0)	5 (31.0)
20.27	Flavoring extracts and syrups	32 (20.8)	11 (27.3)	6 (30.2)
36.26	Primary batteries	–	23 (23.1)	7 (29.6)
36.11	Household vacuum cleaners	27 (22.0)	42 (20.1)	10 (25.6)
All Industries—Median		(12.1)	(12.1)	(13.4)
Top Quintile Bound		47 (18.0)	47 (19.0)	47 (19.1)

Source: U.S. Federal Trade Commission.

[a] Quintile leavers—see text for full information.

[b] Percentages are shown in parentheses.

that accounting data at the line of business and company levels contain strong signals to which important economic decisions are related.

B. *Industry-Level Profitability Changes*

Benston's assertion that year-to-year variations in reported profitability, aggregated to the industry level, are implausibly high (p. 52) can also be confronted with external evidence. Table 3 lists the 21 industries that were among the top 10 industries in terms of operating income-assets ratios in at least one of the reporting years 1974, 1975, and 1976. It reveals some dramatic changes in profitability rank (out of 234 industries for 1974 and 1976 and 237 for 1975). The question Benston poses but fails to answer is, do those moves correspond to "changes in the environment," or must one accept Benston's more negative view that sample changes or measurement errors were to blame? We focus here only on the 11 industries that moved out of the most profitable industry quintile in one or more of the years 1974–76. Those were, to repeat, years of extraordinary turbulence, beginning under price controls (until April 30, 1974), progressing into a price explosion and then a sharp recession, followed in 1976 by a strong recovery.

Industry 33.04. Annual average New York lead prices rose 37 percent from 16.3 cents per pound in 1973 to a record 22.5 cents per pound in 1974.[12] They fell slightly to 21.5 cents in 1975 and rose to 23.1 cents

[12] Unless otherwise indicated, all price information is drawn from various issues of the U.S. Bureau of the Census *Statistical Abstract of the United States* and from the U.S. Department of Agriculture annual, *Agricultural Statistics.*

in 1976. The pattern is fully consistent with the pattern of profitability changes.

Industry 20.17. When price controls were relaxed, the distortions they and sugar import quotas had caused led to a price explosion. Refined sugar prices (at New York) soared from 13.8 cents per pound in 1973 to 33.7 cents in 1974. Refiners had entered fixed proportion revenue-sharing contracts with growers, and so as refined sugar prices climbed, refiner profits rose sharply. See Keith Anderson et al. (1975, pp. 67–76). With expanded output in 1975, prices, like profits, dropped modestly to 30.8 cents. In 1976, following a relaxation of sugar import quotas and greatly increased imports, prices fell sharply to 18.85 cents. The evidence is completely consistent, although we supplement it by jumping out of order to a related industry.

Industry 20.13. In the early 1970's, the wet corn milling industry introduced a new product, high fructose corn syrup (HFCS), which is a close substitute for cane and beet sugar in many applications. For the rest of the story, we can do no better than quote Michael Porter and Michael Spence:

> Demand growth was aided by a tremendous surge in sugar prices in 1974.... However, at the end of 1974 the sugar price support legislation in the United States lapsed and was not renewed because of high sugar prices. The latter then tumbled to the eight cent per pound level. This adversely affected the HFCS market. By 1976 the capacity planned in 1973–74 was coming on-stream, while demand was falling off. By late 1976 industry capacity utilization was low...and the profits had been squeezed out of the margins in the industry. [1982, p. 281]

Industry 28.14. Price controls also distorted fertilizer supply, while sharply higher grain prices increased plantings and hence the derived demand for fertilizers. See Marvin Kosters (1975, pp. 85–89), and Milton David et al. (1976). Fertilizer makers promised capacity expansions and diversion of output from export to domestic markets in exchange for early relief from price controls. At first, prices soared. But when the new capacity came on stream, prices declined, as the two representative price series reported in Table 4 show. Meanwhile the costs of important inputs, especially natural gas, were rising rapidly. The mild and then sudden drop in fertilizer industry profitability is consistent with these movements in output, cost, and prices.

TABLE 4—MOVEMENTS IN THE PRICE PER TON OF REPRESENTATIVE FERTILIZERS

Year	46 Percent Superphosphate	Ammonium Nitrate
1973	$ 91	$ 74
1974	169	155
1975	197	171
1976	152	136

Source: U.S. Department of Agriculture.

Industry 28.15. Price and output behavior is more heterogeneous for the complex mix of pesticides, herbicides, and other agricultural chemicals produced by this industry. The most that can be said is that the profitability changes are similar to those in the closely related fertilizers category, but strong product differentiation inhibited a sharper decline as capacity caught up to demand.

Industry 29.03. Any comprehensive industry reporting system must have some "miscellaneous" or "catch-all" categories, and this is one. Its products include lubricating oils blended outside refineries, petroleum coke, charcoal briquettes, and brake fluids. Only 5 or 6 companies reported in any given year, and we do not know (or would not be permitted to disclose) what products they emphasized, and hence what caused their profitability changes.

Industry 26.01. The producer price index for wood pulp rose from 128.3 in 1973 to 217.8 in 1974, 283.4 in 1975, and 286.0 in 1976. Falling profits despite high and (until 1976) rising prices can be attributed in part to a drop in capacity utilization from above 90 percent in 1974 to less than 85 percent in early 1975 (Harbridge House, 1976, p. 96). In addition, 22 percent of the U.S. timber supply came from federal forests, where the price of the most important pulp input,

TABLE 5—POULTRY INDUSTRY OUTPUT AND PRICES

	1973	1974	1975	1976
Poultry and Egg Output (1967 = 100)	106	106	103	110
Broiler Output (billion pounds)	11.22	11.32	11.10	12.52
Average Broiler Prices (cents per pound)	24.0	21.5	26.3	23.6
Corn Prices (per bushel)	$2.55	3.02	2.54	2.15

Source: U.S. Department of Agriculture.

stumpage, adjusts to downstream market changes only as multiyear tract leases are rebid. Thus, input costs probably followed pulp prices up only with a lag.

Industry 27.06. The manifold business forms industry supplies a wide array of custom-made forms. No adequate output price index is available. It seems reasonable to suppose that form producers found it difficult to pass along fully the rapidly rising prices of a key input, paper.

Industry 33.05. The average price of prime western zinc soared from 20.7 cents per pound in 1973 to a record 35.9 cents in 1974. The price continued a more gradual rise to 39 cents in recession year 1975, while domestic output fell 21 percent. The reason for price increases despite recession is not clear, but decreased profitability is consistent with significantly reduced capacity utilization and/or upward cost curve shifts. The average price fell to 37 cents in 1976 while output rose to 90 percent of its 1974 level.

Industry 34.09. Fabricated structural metals is a contracting industry, assembling bridge sections, joists, television towers, ship sections, and the like. Having written contracts anticipating something like a continuation of the 5.9 percent 1973 inflation rate on its principal input, structural steel, it was shocked by a 28 percent annual increase when price controls expired in April 1974. Wages also rose rapidly, and many bankruptcies occurred.[13] The high profits in 1975 and 1976 evidently came from some combination of capacity reductions (for example, Bethlehem Steel Corp. closed four fabricating plants) and the projection of high steel price inflation in bidding for new contracts. Materials costs fell from 44 percent of sales in 1974 to 35 percent in 1975. The volatility of industry earnings owing to changes in price-cost spreads is enhanced by low capital intensity. In 1974, fabricated structural metals ranked 206th among 234 industries in the ratio of total assets to sales.

Industry 20.02. The poultry and egg processing industry is clear winner in the "profit roller coaster" competition. Preliminary insight into its volatile profit behavior is gained by examining the four output and price series summarized in Table 5. Broiler chicken prices dropped by 10 percent in 1974, partly because cattle held back until meat price controls were lifted in late 1973 surged into the market, forcing down all meat prices. Meanwhile feed prices rose sharply. These two developments are sufficient to explain the poor 1974 profits.[14] The situation reversed in 1975, in part because of lagged output adjustments (broilers require two months to mature from the chick stage). The resulting profit margins were characterized by the *Wall Street Journal* as record breaking.[15] Another overreaction followed: output rose 13 percent in 1976 and broiler prices fell, confirming the *WSJ*'s July

[13] "No Physical Growth Ahead with 5% Dollar Gain," *Engineering News Record*, January 23, 1975, p. 38.

[14] "Poultry Industry Faces 'Rough Sledding,'" *Feedstuffs*, June 24, 1974, *46*, pp. 1, 64.

[15] "U.S. Chicken Industry Enjoying a Feast as Domestic and Foreign Demand Booms," *Wall Street Journal*, July 6, 1976, p. 28.

1976 prediction that the "gravy days may be ending."

To sum up, for most of the eleven industries that experienced quintile-crossing profit variability over the years 1974–76, it is possible with a modest amount of research to find a plausible explanation consistent with economic theory and contemporary published materials. Benston could have found the relevant materials as easily as we. It is, to be sure, easier to sit ex cathedra and conjure up imaginary sampling biases or data deficiencies. But it is also less constructive. Moreover, to anyone who enjoys economic detective work, the Line of Business data offer interesting challenges that can lead, we believe, to insights much more enlightening than those Benston offers.

IV. The Individual Studies

The lack of care with which Benston undertook his critique is demonstrated by the number of errors he makes. Certain errors have been pointed out already. In the Detailed Appendix Comments, we provide a selective addendum focusing on his critique of our papers. We proceed seriatim, identifying page number and column.

V. Conclusion

Data are fallible. So are scholars. Yet when an article is as consistently negative as Benston's, one suspects bias, and when it contains as many demonstrable errors as Benston's, one suspects a degree of carelessness incompatible with the burden a scholar must bear when he singles others' work out for criticism. We also have a more fundamental objection. It is easy enough to sit at one's desk and take pot shots, accurate or inaccurate, at others' empirical research. It is more difficult to augment an already complex data base and design sensitivity tests to ensure that results are robust and not afflicted by bias. Although constructive criticism is necessary and welcome, knowledge can scarcely advance without an emphasis on data base building, testing, and sensitivity analysis. Much has been done using the Line of Business data base to advance economic knowledge. Much remains to be done. From Benston's selective, inaccurate criticism of our work, readers of this *Review* may have gained the opposite impression. We hope to have corrected it at least in part, but we believe the best proof of our position must come from the reader's own unbiased perusal of the substantial literature emanating from *LB* data.

DETAILED APPENDIX COMMENTS

P. 54–2; Benston is correct in his observation that Mueller's (1980) study illustrates the difficulty in estimating production functions using the existing *LB* data. Unfortunately, the *LB* program did not collect the requisite physical output and employment measures. Mueller concluded that the data were insufficient and abandoned his effort. Where Benston errs is in insisting that without production function estimates, "structure-performance studies cannot be used" (p. 55–1) to illuminate such questions as economies of scale. Compare Ravenscraft (1984) and Long (1982), who develop techniques by which scale economy effects can be disentangled.

P. 55–1 (fn. 37, and also p. 58, fn. 43): A stepwise procedure was used by Ravenscraft not to eliminate variables from the analysis, as alleged by Benston, but to select variables for a heteroskedasticity correction.

P. 55–2 (fn. 38): Contrary to the implication drawn, Martin found significant differences between profit/sales and profit/assets regressions mainly when nontraceable assets were excluded from the denominator. And the Herfindahl index was significant in two cases, not one as alleged.

P. 56–1 (and also p. 59–1): Benston correctly notes that the negative and significant *LB* assets/sales coefficient was inconsistent with a priori expectations. However, accounting bias is not the only explanation. Ravenscraft (1983) shows that industry assets/sales and *LB* assets/sales, the latter interacting with market shares, have the anticipated positive signs. Thus, the negative coefficients on *LB* assets/sales imply low profits for low market share lines with high asset intensity.

P. 56–2: Citing Stanley Ornstein (1975), Benston argues that "the relationship between profit/sales and concentration (or market share) should be specified as log linear...." Ornstein's argument, however, is not based upon a profit-maximizing model, and it ignores the fact that profits can be meaningfully negative. Roger Clarke and Stephen Davies (1982) and Long (1982) use a standard oligopoly model to derive first-order equations in which profits/sales is a *linear interactive* function of concentration and market share. This form, rejected by Benston despite strong theoretical support, was estimated by Long (1982), Ravenscraft (1983, Table 2), and Kwoka and Ravenscraft.

P. 60–2: Scott's analysis covered 246 industry categories, not the "twenty-four lines" claimed by Benston.

P. 61–2: Benston claims that Scherer's methodological paper did not relate *R&D* expenditures to productivity. In fact, the published version cited in Benston's references did report such relationships.

P. 62–1: Benston claims that it "seems obvious that if more is spent on research and development, more patents will be applied for and granted." The earlier FTC version of his paper (1982, p. 79) contains nearly the same language and also a statement that "*R&D* expenditures should not be expected to relate materially to the number of patents granted" (p. 8). Obviously, something was not obvious. In fact, the paper's main claim to novelty was not that patents and *R&D* were related, but that the relationship was preponderantly linear once a modest size threshold is exceeded, and that other variables have little incremental explanatory power.

P. 62–2: Scherer's productivity growth analysis used three different productivity measures, not two. In the main analysis, the level of aggregation was 3- or 4-digit, not 2-digit. In most cases, the *R&D* outlays were divided by value of output, not value-added. No variable measuring changes in sales/labor expense was used, contrary to Benston's statement. Capital/labor variables were not omitted as alleged, nor did the bulk of the analysis rely upon simple correlations as alleged. A "wrong lag" analysis tested for timing effects, contradicting Benston's conjecture that greater productivity might have led to more intensive *R&D*.

P. 63–1: Long's analysis related patent/sales, not patents, to *R&D*/sales with quadratic terms. Benston's error here is important, since a negative quadratic coefficient sign is not an indication of poor data quality, as suggested, but of plausible diminishing marginal returns in the input-output relationship. There is no information in Long's paper from which Benston could draw the conclusion that the estimated relationships differ in sign between years. Rather, his conclusion reflects a misreading of differences in statistical significance as differences in sign.

REFERENCES

Anderson, Keith B., Lynch, Michael and Ogur, Jonathan, *The U.S. Sugar Industry*, Staff Report, Federal Trade Commission, Washington, July 1975.

Benston, George J., "An Analysis of the Usefulness of the Federal Trade Commission's Line of Business Reporting Program," in *Benefits and Costs of the Federal Trade Commission's Line of Business Program*, Vol. II, *Public Comments*, Washington, September 1982, 267–370.

______, "The Validity of Profits-Structure Studies with Particular Reference to the FTC's Line of Business Data," *American Economic Review*, March 1985, *75*, 37–67.

Benvignati, Anita M., "Domestic Profit Advantages of Multinational Firms," *Journal of Business*, forthcoming 1987.

Clarke, Darral G., "Econometric Measurement of the Duration of Advertising Effect on Sales," *Journal of Marketing Research*, November 1976, *13*, 345–57.

Clarke, Roger and Davies, Stephen, "Market Structure and Price-Cost Margins," *Economica*, August 1982, *49*, 277–87.

David, Milton L. et al., "Study of Causes, Adjustments, and Impacts of Shortages in Fertilizer," in *The Commodity Shortages of 1973-1974*: *Case Studies*, Washington: National Commission on Supplies and Shortages, August 1976.

Demsetz, Harold, "Two Systems of Belief About Monopoly," in Harvey J. Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown, 1974, 164–84.

Friedman, Milton, "The Methodology of Positive Economics," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953, 3–43.

Gale, Bradley T. and Branch, Ben S., "Concentration versus Market Share: Which Determines Performance and Why Does It Matter?," *Antitrust Bulletin*, Spring 1982, *27*, 83–105.

Hagerman, Robert and Zmijewski, Mark E., "Some Economic Determinants of Accounting Policy Choice," *Journal of Accounting and Economics*, August 1979, *1*, 141–61.

______ **and** ______, "A Test of Accounting Bias and Market Structure: Some Additional Evidence," *Review of Business and Economic Research*, Fall 1981, *17*, 84–88.

Kosters, Marvin H., *Controls and Inflation* Washington: American Enterprise Institute, 1975.

Kwoka, John E. and Ravenscraft, David J., "Cooperation vs. Rivalry: Price-Cost Margins by Line of Business," *Economica*, August 1986, *53*, 351–63.

Leamer, Edward E., "Sensitivity Analysis Would Help," *American Economic Review*, June 1985, *75*, 308–13.

Long, William F., "Impact of Alternative Allocation Procedures on Econometric Studies of Structure and Performance," manuscript, Federal Trade Commission, July 1981.

______, "Market Share, Concentration and

Profits: Intra-Industry and Inter-Industry Evidence," manuscript, Federal Trade Commission, 1982.

______, **and Ravenscraft, David F.**, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 494–500.

______ **et al.**, *Benefits and Costs of the Federal Trade Commission's Line of Business Program*, Vol. I, *Staff Analysis*, Washington: Federal Trade Commission, September 1982.

Marshall, Cheri T., "PIMS and the FTC Line-of-Business Data: A Comparison," unpublished doctoral dissertation, Harvard University Graduate School of Business Administration, 1986.

Martin, Stephen, *Market, Firm, and Economic Performance: An Empirical Analysis*, Monograph 1983-1, New York University Graduate School of Business Administration Salomon Brothers Center, 1983.

Mautz, Robert and Skousen, Fred, "Common Cost Allocation in Diversified Companies," *Financial Executive*, June 1968, *36*, 15–25.

Mueller, Dennis C., "Economies of Scale, Concentration, and Collusion," manuscript, Federal Trade Commission, September 1980.

Ornstein, Stanley I., "Empirical Uses of the Price-Cost Margin," *Journal of Industrial Economics*, December 1975, *24*, 105–17.

Porter, Michael E. and Spence, A. Michael, "The Capacity Expansion Process in a Growing Oligopoly: The Case of Wet Corn Milling," in J. J. McCall, ed., *Economics of Information and Uncertainty*, Chicago: University of Chicago Press, 1982, 259–309.

Ravenscraft, David J., "Intracompany Transfer Pricing and Profitability," manuscript, Federal Trade Commission, December 1981.

______, "Structure-Profit Relationships at the Line of Business and Industry Levels," *Review of Economics and Statistics*, February 1983, *65*, 22–31.

______, "Collusion vs. Superiority: A Monte Carlo Analysis," *International Journal of Industrial Organization*, December 1984, *2*, 385–402.

______ **and Scherer, F. M.**, (1986a) "The Profitability of Mergers," Working Paper No. 136, Federal Trade Commission Bureau of Economics, January 1986.

______ **and** ______, (1986b) "Mergers and Managerial Performance," Working Paper No. 137, Federal Trade Commission Bureau of Economics, January 1986.

______ **and** ______, *Mergers, Sell-offs, and Economic Efficiency*, Washington: The Brookings Institution, 1987.

Schmalensee, Richard, "Do Markets Differ Much?," *American Economic Review*, June 1985, *75*, 341–51.

Scott, John T., "Multimarket Contact and Economic Performance," *Review of Economics and Statistics*, August 1982, *64*, 368–75.

______ **and Pascoe, George**, "Capital Cost and Profitability," *International Journal of Industrial Organization*, September 1984, *2*, 217–33.

______ **and** ______, "Beyond Firm and Industry Effects on Profitability in Imperfect Markets," *Review of Economics and Statistics*, May 1986, *68*, 284–92.

Weiss, Leonard W., "The Concentration-Profits Relationship and Antitrust," in Harvey J. Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown, 1974, 184–223.

Harbridge House, Inc., "Spot Shortage Conditions in 1973–1974: The Pulp and Paper Industry Experience," in *The Commodity Shortages of 1973–1974: Case Studies*, Washington: National Commission on Supplies and Shortages, August 1986.

U.S. Bureau of the Census, *Statistical Abstract of the United States*, Washington: USGPO, various years.

U.S. Department of Agriculture, *Agricultural Statistics*, Washington: annually.

U.S. Federal Trade Commission, *Statistical Report: Annual Line of Business Report: 1974–1977*, Washington: USGPO, 1981–85.

The Validity of Studies with Line of Business Data: Reply

By GEORGE J. BENSTON*

I welcome the present extensive comment (F. M. Scherer et al., 1987) on my paper (1985) because the authors present some new analyses and provide us with an opportunity to discuss some important issues. In this reply, I attempt to follow the order of their comment in responding to their criticisms.

I

In their Section I, Scherer et al. emphasize what they believe is a significant advance in knowledge about the previously assumed-to-have-been-proven positive relationship between industry concentration ratios and economic profits. Studies using the *LB* data now show, they say, that Harold Demsetz (1974) rather than Leonard Weiss (1974) was correct—"when leading sellers have unit cost advantages over smaller firms, regression analyses at the industry level could, because of aggregation biases, yield positive profit-concentration coefficient even when no price-raising effect was present" (pp. 205–206). This research, they say,

> has shown that individual market share effects are indeed much more powerful than the traditionally emphasized concentration effects in explaining profitability. With most specifications, concentration coefficients turn out to be *negative*, not positive, in conjunction with market share variables. The positive and significant market share relationships alone cannot discriminate between monopoly power and efficiency or cost advantage hypotheses. [p. 206]

I would have preferred that the commentators had recognized that the positive *reported* (accounting) profits-concentration relationship might have been due to biases (with respect to economic market values) in the profits numbers used in the studies. Instead, they now believe that higher market share but not concentration results in higher economic profits, rather than that collusion reflected in concentration ratios results in higher economic profits. The latter conclusion gave rise to a national antitrust policy mistakingly based on preventing mergers that result in significantly higher concentration ratios. Abandoning this basis for policy would be an advance. But policymakers would be ill-advised to conclude the opposite from these studies—that is, that higher economic profits are due to individual companies' market shares rather than to collusion. It may be, as the authors point out (while missing my point), that "firms in more concentrated markets tended to choose accounting policies that *reduced* their reported profits" (p. 206). As I show in my original article, the biases in the accounting numbers can go either way and could wash out in the aggregate, thereby obscuring important relationships. In addition, as Betty Bock (1975), Weiss (pp. 194–96), and others have shown, 4-digit SIC-defined "markets" conform very poorly to the economic definition of markets. Therefore, I still must conclude that these studies do not provide useful evidence on either side of the debate. All the new studies appear to show is that concentration ratios derived from SIC-defined data are not meaningful measures of the potential for collusion.

Nor do the new studies reviewed in the comment (Section II, Part A) "show that [*economic*] profitability doubled or even trebled with increases in market share over the

*Professor of Accounting, Economics and Finance, William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627. The paper preceding my 1985 paper, which was submitted to the Federal Trade Commission, was supported financially by the Business Roundtable. The Roundtable did not support or review the considerable rewriting of the paper submitted to the *American Economic Review*; hence, their prior support was not gratefully acknowledged. I take this opportunity to recognize their help.

range of observed values" (p. 207). Rather, as I attempted to point out and as the commentators seem not to recognize, the relationship measured could be due plausibly to the way accounting profits are recorded. Consider, for example, a firm that plans to dominate the market for a product. It engages in an extensive research and development program in years 1 and 2. The costs of this program are written off as current-year expenses—no asset is recorded. If the program is effective, new and/or improved products are developed. Changes in manufacturing processes and training of production and sales personnel are accomplished in years 2 and 3. These costs are written off. The product is advertised and promoted in years 3 and 4, the costs of which are written off. In years 4 and 5, the product is accepted by the public and the company's market share increases. Higher profits are now recorded, in large part because the costs of researching, developing, efficiently producing, and promoting the product were charged to expense in previous years. An *ex post* present-value analysis might show that the market-domination plan reduced the shareholders' wealth. (*Ex ante* the plan is likely to have been judged an expected positive-present-value project.) I suggest that this accounting-bias explanation for the "robust" market share-accounting profits regression coefficients presented by Scherer et al. (Table 1) is preferable to the explanation they imply—that firms with higher market shares are more efficient, although both explanations are plausible. But even if the efficiency explanation were correct, one cannot infer that higher market shares result in greater efficiency—the reverse causation is equally likely. Indeed, the commentators incorrectly assert causality in several places. For example, they state that a study "*reveals* that the profit advantage of larger sellers *depends* upon structural variables which in turn *influence* how the market leader *chooses* to exploit its strategic advantages" and "Analyses...show the market share-profit relationships to *follow* from a mixture of cost, first mover, and subjectively determined quality advantages enjoyed by larger sellers" (p. 206, emphasis added).

Scherer et al. point to the privately financed PIMS (Profit Impact of Market Strategies) program as showing that "the kind of data collection Benston attacks as meaningless has passed a clear market test" (p. 209). But I did not claim that accounting data were useless. To the contrary, my paper includes a subsection (p. 50, II, Part B.2) that describes how such data are used by managers who know the environment that gave rise to the data. Presumably, those companies that participate in the PIMS program (which, unlike the FTC *LB* program does not use the SIC descriptions to define "lines of business") find those data useful for some purposes. But such was not the case for the enterprises that were required to report their numbers to the FTC. Indeed, virtually none of the respondents accepted the program and many took legal action to block it.

In Section II, Part B, of their comment, Scherer et al. consider three "other accounting impact analyses." They first criticize my having pointed out that a random allocation of nonallocable (i.e., not causally related) overhead to lines of business is as valid as an allocation based on direct expense or sales because there is no conceptually correct way to allocate such costs. The fact that allocations based on direct expense or sales appears to be "more realistic" does not result in numbers that reflect economic market values.[1]

Scherer et al. next claim that, in reviewing David Ravenscraft's (1981) work on the effect of transfer pricing, I did not recognize that a numerator "divided by zero or something close to zero leads to very large numbers of the sort Ravenscraft reported[9]" (p. 210). Remarkably, that possibility did not escape my purview. Contrary to the implication of their footnote 9, I not only cited the study by Robert Mautz and Fred Skousen (1968), but also reported the effect of eliminating

[1] I do not understand why allocating overhead in accordance with direct expenses or sales results in numbers that are "competitively sustainable" with the exception of cost-reimburse contracts. I do not believe that I misinterpreted William Long's (1981) paper.

observations in their data with very low denominator values.

Finally, the commentators point to unpublished studies by Ravenscraft and Scherer (1986) that "show that [when] postmerger profit rates of acquisition-prone lines of business...[have] substantial amounts of purchase accounting assets...[they] report significantly lower profitability postmerger than lines using pooling of interests accounting" (Scherer et al., p. 210). It is good that they recognize this rather straightforward effect of the choice of accounting method on reported profits. I wish they had done more of this type of "analysis," or at least explicitly recognized this and other effects of accounting method on reported profits.

II

In their Section III, Part A, Scherer et al. present a summary of an analysis indicating that lines of business with negative returns tend to be sold off. I applaud this work (the details of which I have not seen) and hope that it includes an analysis of lines of business with relatively high negative returns that were not sold off as well as lines with relatively high returns that were sold. In considering the effect of accounting-measured profitability on the managers' decision to disinvest, I would hope that the researchers can distinguish between the hypothesis that the accounting data give meaningful signals of economic returns and the hypothesis that the managers are attempting to increase earnings per share at the expense of making net present value reducing choices. I also hope that the authors consider the possibility that some of the returns are due to accounting practices.

I also applaud the analysis presented in their Section III, Part B, although it would be presumptuous of me to assume that they had followed my suggestion that they do detailed company and industry studies (see my original paper, p. 65). If descriptions of the type they illustrate were done for all the data, the models might be better specified or it might be recognized that models using large-scale heterogeneous data sets cannot be adequately specified. Indeed, had he done this type of analysis, Scherer might not have drawn the following conclusion at the Georgetown Private Antitrust Litigation Conference in November 1985. He first explained: "I served as economic advisor to Judges Robson and Will in the Folding Carton Case (MDL 250). At the time damage claims were under consideration, we sought systematic evidence on how the cartel's breakup in 1974 affected industry profitability" (1987, p. 2). He then presented a table showing operating income as a percent of assets (*OI/A*) for FTC industry category 26.10 (paperboard containers and boxes). For 1974 through 1977, these numbers are 17.4, 11.7, 9.3, and 5.0 percent, from which Scherer concluded: "[t]he share fall in returns with the transition from conspiracy to competition in paperboard containers is unmistakable" (p. 3). Though Scherer et al. note the effect on recorded sales prices of the removal of price controls in 1974, Scherer (1987) did not mention the possibility that the relatively high 1974 *OI/A* might have been due to this factor rather than to a conspiracy. He also did not consider the possibility that 1974 was a random event. As a means of checking this possibility, I obtained the following Net Profit/Sales (before tax) (*NP/S*) figures for folding cartons from the Meade Corporation. These numbers show that during the presumed conspiracy *NP/S* varied from 0.3 to 5.4 percent between 1960 and 1968. From 1968 through 1971 they were between 10.9 and 12.1 percent. In 1972 and 1973, *NP/S* was 6.9 and 7.2 percent, after which the percentage increased to 12.7 percent in 1974 and then decreased in 1975, 1976, and 1977 to 8.6, 7.1, and 7.8 percent. If there were a conspiracy between 1960 and 1975, as Scherer believes, these numbers do not reflect it well.

III

Scherer et al. charge me with making "numerous errors" in characterizing their work. Given this, I must reply in my Detailed Appendix Comments to each of the eleven "errors" they list. (The references are to the pages in my original paper given in their comment.) Of the purported errors, I believe that only those labeled pages 55–1, 60–2, 55–2, 62, 63–1, and small portions of their footnote 5 are really errors, and these consist of minor misdescriptions of specifics about the authors' empirical work.

IV

I believe that a review of their papers and of my responses to their present and previous allegations of my having made errors shows that they are no less guilty than am I of a fall from perfection. Indeed, considering that they wrote the papers and had the advantage of combining the talents of eight coauthors, it is remarkable that they made more errors in characterizing their work than I did.

What is important is their failure to answer the basic thrust of my paper—that the FTC's *LB* program, which imposed considerable costs on the respondents, was undertaken without sufficient hypothesis development, including an adequate consideration on whether and how the accounting biases in the data could be overcome. I do not disagree with the commentators' admonition that "knowledge can scarcely advance without an emphasis on data base building, testing, and sensitivity analysis" (p. 215). But I suggest that brute empiricism cannot overcome the basic limitation inherents in the studies I reviewed in my paper and in a large number of similar profits-structure studies, and that accounting profit/sales or profit/assets is not a valid measure of economic profitability.[2] Perhaps some day researchers such as Scherer et al. will be able to "solve" this problem. I believe that some portions of their comment represent an advance. However, they have not shown that they are as yet able to use such data as were gathered by the FTC's *LB* program to draw meaningful conclusions about economic market relationships.

DETAILED APPENDIX COMMENTS

Footnote 5: I agree; the year headings are reversed and the negative sign on the market-share coefficient in the Weiss-Pascoe regression should be positive. I did adjust the coefficients to account for differences in the scale of the independent variables used by Ravenscraft and Stephen Martin (1983). However, I made two errors. The Ravenscraft coefficient of the buyer concentration index should be −3.14 rather than −31.40 (compared to −.09 for Martin), and the coefficient of Martin's buyer dispersion index should be 6.13 rather than −6.13 (compared to −.46 for Ravenscraft). Thus all the signs of their variables are not equal. These authors did not identify these errors either in the comment or in response to previous correspondence. Rather they continue to claim falsely that I did not adjust the data for differences in scale, despite the statement in my original paper that Ravenscraft's coefficients were "[r]escaled to Martin's magnitudes." (fn. b, Table 3, p. 57, and also on p. 59.) With respect to the comparison between Scott and Weiss-Pascoe, the notes to my Table 3, and footnote 47 state that "[d]ifferences in the coefficients' magnitudes could be due to the scaling of the variables; [an adjustment could not be made because] the means are not reported" (p. 61).

P. 54–2. I did not write that "without production function estimates, 'structure-performance studies cannot be used' to illuminate such questions as economies of scale" (Scherer et al., p. 215). I wrote: "the *LB* data are not even amenable to estimating what might be labeled (incorrectly) as economies of scale, without which the structure-performance studies cannot be used for policy purposes" (p. 55). Furthermore, even if Dennis Mueller (1980) conceptually were able to estimate cost functions and found economies of scale, he could not have distinguished between the hypotheses that firms dominate markets because they are more efficient (have lower actual costs), or have lower reported costs because they previously expensed assets that gave them market dominance, or have lower reported costs because of other biases in the accounting numbers. Nor, as I pointed out in my original paper, could he determine "whether the regressions trace out cost curves, demand curves, or the intersection of individual demand and cost curves" (p. 54–2). Ravenscraft (1984) does not use the FTC *LB* data; hence, whatever its merits, it is not relevant to my criticism. Perhaps I did not understand what Long (1982) did. In any event, I see no reason for the authors to state that I erred.

P. 55–1: Mea culpa. I misread Ravenscraft's paper.

P. 55–2: I believe that my statement "[t]he coefficients differ considerably in significant and sometimes in sign" is not incorrect. Nontraceable assets were excluded from six of Martin's twelve regressions. I compared the other three pairs of identically structured profit/sales and profit/assets regressions, which the commentators imply have similar coefficients. Thirty percent of the 39 coefficients are significant at the .05 or below level in one regression, but not significant at the .10 or below level in the comparable regression. Of these, 16 percent had different signs in each regression. An additional 3 percent had significant coefficients in

[2] Skeptical readers should be convinced by Franklin Fisher and John McGowan (1983), the comments and discussion on their article published in the June 1984 issue of this *Review*, and Fisher (1986).

both regressions but of different signs. However, I incorrectly state that the Herfindahl index was significant in one case when it was significant in two of six regressions.

P. 56–1: First, I never claimed that "accounting bias is... the *only* explanation" of any result (emphasis added), which is not to say that the commentators' explanation is satisfactory. Second, this hardly appears to qualify as factual error.

P. 56–2: I suggest that readers who might believe that a linear regression of profit/sales yields results that are meaningful for understanding economic relationships and for antitrust policy should read Stanley Ornstein (1975) and Stanley Liebowitz (1982, 1984), and perhaps the discussion in my original paper, pp. 55–57.

P. 60–2: Mea culpa—John Scott (1982) covered 246 rather than 24 industry categories.

P. 61–2: A regression of one variable on another is a "relationship" of sorts, but not one that provides a meaningful answer to the question that Scherer states motivated the study: "what quantitative *links* exist between *R&D* and productivity growth... "? (Scherer, 1981a, cited in my original paper, pp. 61–62, emphasis added).

P. 62–1: What seems obvious (to me) is that it is not very useful to learn the precise relationship that one might be able to measure between expenditures on *R&D* and the number of patents granted. Over a wide-ranging scale of activities, it would be difficult to believe that, say, a $50 million expenditure on *R&D* would not result in more patents granted than a $50 thousand expenditure on *R&D*. Nor would a linear relationship be surprising or interesting. However, any given amount of expenditures might easily give rise to a larger or smaller number of patents, such that any prediction of the number of patents that would be forthcoming from a given expenditure would have a large error. In any event, they have not answered the doubt raised in my original paper—what could be the possible value of an empirical measurement of relationships such as those presented by Scherer? Can such data provide "information of technological change," which is what he purported to measure?

P. 62: In my review of Scherer's and the others' papers, I could not possibly describe and discuss every regression that they ran without duplicating their work. In this portion of my paper I was referring to the regressions presented in Scherer's Table 1 (1982). In that table he uses only two productivity variables. (In any event, I could not find the third productivity measure after rereading the paper.) The data are described by Scherer as "2-digit manufacturing group results" (p. 629). (I still cannot find where Scherer states that he used 3- or 4-digit aggregation levels, and the commentators did not respond to my request for a specific reference.) A footnote to Scherer's Table 1 explains: "All *R&D* variables are divided by 1974 industry value added." With respect to the definition of a dependent variable, "total factor productivity growth," I believe that an empiricist should fully describe his or her data and not rely on the readers understanding the definition the researcher has in mind. The capital-labor variables were not included in Table 2. This set of regressions follows the description of the model. However, I agree that the variable was included in later regressions and that such regressions were presented in addition to the simple correlations that I mentioned. But I do not agree that Scherer's "wrong lag" analysis demonstrates the timing effect claimed for it, as I tried to explain in my paper.

P. 63–1 [63–2]: Mea culpa—I should have said that the number of patents issued/sales was regressed on *R&D* expenditures/sales and its square (W^2). My error might have been due to the variable in question having been identified on page 1 of Long's (1980) paper and used empirically on page 10. Long reports finding "57 of the 203 industries with coefficients of W^2 which were significantly different from zero. Of these, 30 coefficients were positive and 27 were negative..." (p.10). I read this statement as saying there were differences in sign. Actually, I had intended to refer to Long's first set of regressions of *R&D*/sales on market share. The coefficients of market share differ in sign within and between years. But none of this is important. The question remains: what can the *LB* numbers tell us about the purpose of the paper, which Long says, is "the determination of research and development expenditures for the improvement of the quality of manufactured products"?

REFERENCES

Benston, George J., "The Validity of Profits-Structure Studies with Particular Reference to the FTC's Line of Business Data," *American Economic Review*, March 1985, *75*, 37–67.

Bock, Betty, "Line of Business Reporting: A Quest for a Snark?," *The Conference Board Record*, November 1975, *12*, 10–19.

Demsetz, Harold, "Two Systems of Belief About Monopoly," in Harvey Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown, 1974, 164–84.

Fisher, Franklin M., "On the Misuse of the Profits-Sales Ratio to Infer Monopoly Power," manuscript (rev.), Massachusetts Institute of Technology, 1986.

______ **and McGowan, John J.**, "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

Liebowitz, Stanley J., "What Do Census Price-Cost Margins Measure?," *Journal of Law and Economics*, October 1982, *25*, 231–46.

______, "On the Measurement of Monopoly Power," manuscript, University of

Rochester, 1984.

Long, William F., "Advertising Intensity, Market Share, Concentration and Degree of Cooperation," manuscript, Federal Trade Commission, September 1980.

______, "Impact of Alternative Allocation Procedures on Econometric Studies of Structure and Performance," manuscript, Federal Trade Commission, July 1981.

Martin, Stephen, *Market, Firm, and Economic Performance: An Empirical Analysis*, Monograph 1983-1, New York University Graduate School of Business Administration Salomon Brothers Center, 1983.

Mautz, Robert and Skousen, Fred, "Common Cost Allocation in Diversified Companies," *Financial Executive*, June 1968, *36*, 15–17, 19–25.

Mueller, Dennis C., "Economies of Scale, Concentration, and Collusion," manuscript, Federal Trade Commission, September 1980.

Ornstein, Stanley I., "Empirical Uses of the Price-Cost Margin," *Journal of Industrial Economics*, December 1975, *24*, 105–17.

Ravenscraft, David J., "Intracompany Transfer Pricing and Profitability," manuscript, Federal Trade Commission, December 1981.

______, "Collusion vs. Superiority: A Monte Carlo Analysis," *International Journal of Industrial Organization*, December 1984, *2*, 385–402.

______ **and Scherer, F. M.,** (1986a) "The Profitability of Mergers," Working Paper No. 136, Federal Trade Commission Bureau of Economics, January 1986.

______ **and** ______, (1986b) "Mergers and Managerial Performance," Working Paper No. 137, Federal Trade Commission Bureau of Economics, January 1986.

Scherer, F. M., "Comment on the Salop-White and Teplitz Papers," in Lawrence J. White, ed., *Private Antitrust Litigation: New Evidence, New Learning*; Boston: MIT Press, 1987.

______, "Inter-Industry Technology Flows and Productivity Growth," *Review of Economics and Statistics*, November 1982, *64*, 627–34.

______ **et al.,** "The Validity of Studies With Line of Business Data: Comment," *American Economic Review*, March 1987, *77*, 205–217.

Scott, John T., "Multimarket Contact and Economic Performance," *Review of Economics and Statistics*, August 1982, *64*, 368–75.

Weiss, Leonard W., "The Concentration-Profits Relationship in Antitrust," in Harvey Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown, 1974, 184–233.

Credit Rationing: A Further Remark

By John G. Riley*

The recent, widely cited papers on credit rationing by Joseph Stiglitz and Andrew Weiss (1981, 1983) show that, in the presence of asymmetric information, it is quite possible for credit rationing to occur even with unregulated, competitive banking. More precisely, they show that circumstances can arise in which not all those in a pool of observationally equivalent loan applicants will be offered loans.

Stiglitz and Weiss (S-W) consider the potential for credit rationing in a single pool of loan applicants, each of whom have projects with the same expected return. Here the focus is shifted to the banking sector's total demand for loanable funds—a demand derived by examining the effects of interest rate changes across different risk pools. It is shown that even if adverse selection occurs in each risk pool, only in a single marginal pool could rationing ever be observed. From this it is concluded that the extent of rationing generated by the S-W model is not likely to be empirically important.

The structure of the S-W model is as follows. Banks identify a pool of loan applicants P_μ who have projects with equal borrowing requirements $\bar{L}$ and equal expected gross returns μ.

Banks are also aware that, while projects have identical mean returns, they differ in their riskiness. A loan applicant in risk class s has a cumulative distribution function for the gross return x of $F(x-\mu; s)$. For simplicity, it is assumed that loan riskiness can be ranked by the second-order stochastic dominance criterion and that a higher s represents a more risky loan.[1] Informational asymmetry is introduced by assuming that each loan applicant knows his risk class s, but lenders know only that the loan applicants are distributed according to some underlying cumulative distribution function $G(s)$.

In the absence of collateral possibilities, the return to a loan applicant in risk class s, given a gross interest rate r, is

$$(1) \qquad A_s(x) = \mathrm{Max}\{x - r\bar{L}, 0\}.$$

If a lender must pay an interest rate i on loanable funds, a loan to an applicant in risk class s yields

$$(2) \qquad B_s(x) = \mathrm{Min}\{r\bar{L}, x\} - i\bar{L}.$$

The key to the Stiglitz and Weiss analysis is that an applicant's "utility" function $A_s(x)$ is zero for sufficiently low returns and linear and increasing for x exceeding $r\bar{L}$. That is, $A_s(x)$ is a *convex* function of the return x. Therefore, if two projects have the same mean, it is the more risky which generates a higher expected gain to an applicant and thus a lower expected gain to the lender. It follows that, as the interest rate facing loan applicants rises, it is the less risky projects which are the first to have an expected return below the loan applicants' reservation income level $\bar{A}$. Raising the interest rate therefore results in adverse selection.

While a higher lending rate r raises a lender's gross expected return on a given subset of borrowers, the adverse-selection effect tends to lower the gross expected return. Indeed a highly risky opportunity with a very low probability of success will yield a gross return of close to zero to the lender. Thus, for sufficiently high interest rates, the

*University of California, Los Angeles, CA 90024. This paper was completed while I was a visiting lecturer at the University of Canterbury. The helpful suggestions of John Haltiwanger, Hellmuth Milde, Michael Waldman, and a referee are gratefully acknowledged.

[1] Formally,

$$\partial/\partial s \int_0^y F(x-\mu, s)\, dx \geq 0.$$

This condition implies that the variance of x is an increasing function of s, thus one can think intuitively of a "more risky" loan as one having higher variance.

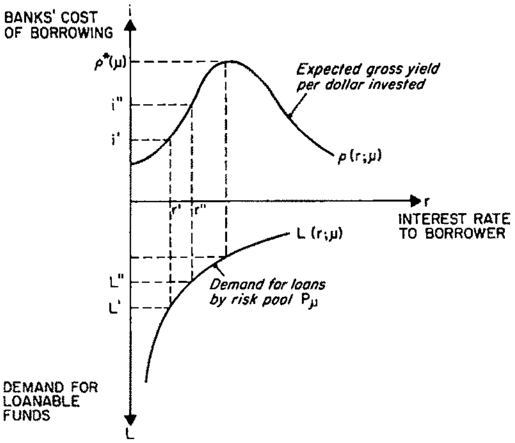

BANKS' COST OF BORROWING
i
$\rho^*(\mu)$
i''
i'
Expected gross yield per dollar invested
$\rho(r;\mu)$
r
INTEREST RATE TO BORROWER
r'
r''
L (r;μ)
L''
L'
Demand for loans by risk pool P_μ
DEMAND FOR LOANABLE FUNDS
L

FIGURE 1. DERIVING THE DEMAND FOR LOANABLE FUNDS

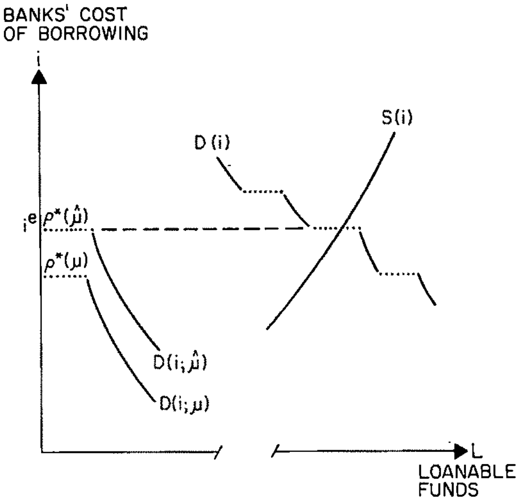


FIGURE 2. EQUILIBRIUM IN THE MARKET FOR LOANABLE FUNDS

gross expected return per dollar invested, $\rho(r;\mu)$, declines with r. Purely for expositional ease, it is assumed, following Stiglitz and Weiss, that $\rho(r;\mu)$ has a unique turning point. The maximum value is denoted by $\rho^*(\mu)$.

Given the assumption that, for each pool of loan applicants, the return x has the same distribution about its mean μ, it follows that $\rho(r;\mu)$ and hence $\rho^*(\mu)$ are strictly increasing functions of μ.[2]

With this as background it is helpful to shift the focus to the demand by banks for loanable funds. Consider Figure 1. In the top half of the figure the gross return $\rho(r;\mu)$ is depicted as a function of the interest rate r. In the bottom half of the figure the loan applicants' total demand for funds is shown as a decreasing function of r, reflecting the exit from the market by preferred risk classes as r increases.

Suppose banks must pay a gross interest rate i' on loanable funds. Competition among banks will drive the lending rate down to the point r' where the expected gross return $\rho(r';\mu)=i'$. But, at this interest rate, the demand by loan applicants and hence the banking industry's derived demand for loanable funds is L'.

Arguing in exactly the same way, at a higher interest cost of loanable funds i'', the competitive interest rate offered to the pool of loan applicants is r'' and the derived demand for loanable funds is L''. Formally, the equilibrium borrowing rate $r=R(i;\mu)$ is given implicitly by the zero profit condition $\rho(i;\mu)=i$. It is therefore possible to graph the entire derived demand for loanable funds:

$$L=D(i;\mu)\equiv L(R(i;\mu);\mu).$$

This is depicted in Figure 2.

Given our assumptions, the derived demand declines continuously as i increases until $i=\rho^*(\mu)$. At this point, the expected return per dollar invested in the pool is at its maximum. Therefore, if the bank borrowing rate rises beyond $\rho^*(\mu)$, the derived demand for loanable funds by the banking industry drops to zero. Thus the demand curve $D(i;\mu)$ has a discontinuity at $i=\rho^*(\mu)$.

The final step in the argument is to aggregate across observably different risk classes. Since the maximum expected yield $\rho^*(\mu)$ increases with μ, the discontinuity oc-

[2] It should be intuitively clear that $\rho^*(\mu)$ is a strictly increasing function under much weaker assumptions as well.

curs at a higher bank borrowing rate for risk pools with a greater mean return. Summing over risk classes yields the banking industry's aggregate demand for loanable funds

$$D(i)=\sum_{k}D(i;\mu_k).$$

This is also shown in the right side of Figure 2, along with a positively sloped supply curve for loanable funds. As depicted, all those pools of loan applicants for whom $\rho^*(\mu)$ is less than the equilibrium cost of loanable funds i^e are unprofitable and are excluded. For these groups, banks are simply unwilling to lend. Equivalently, from the loan applicants' viewpoint, the implicit borrowing rate is so high that even the most risky projects are unprofitable. On the other hand, each risk pool with $\rho^*(\mu)>i^e$ faces an equilibrium borrowing rate $r^e=R(i^e;\mu)$ and, as depicted in Figure 1, supply equals demand.

For none of these groups is there any "rationing" in the normal sense. Figure 2, however, illustrates a case in which the supply curve cuts the demand curve on a horizontal segment. That is, there is one risk pool with mean return $\hat{\mu}$ such that $\rho^*(\hat{\mu})=i^e$. In this risk pool, and only in this risk pool, is there excess demand for loans. The conclusion to be drawn therefore, is that the extent of rationing implied by the S-W model is not likely to be very important empirically.

There is, however, one clear empirically testable implication. For each risk pool there is an equilibrium interest rate $r^e=R(i^e;\mu)$. It is readily confirmed that this interest rate strictly increases with the equilibrium cost of loanable funds i^e and decreases with the mean return of the risk pool, μ. Moreover, and this is the critical point, there is some maximum interest rate $r(i^e,\hat{\mu})$ at which any lending takes place. All loan applicants in pools with mean returns below $\hat{\mu}$ face discontinuously larger interest rates and loan demand drops to zero.

It is interesting to consider how such a market responds to a supply shock. Suppose there is an upward shift in the supply curve $S(i)$. From Figure 2, the equilibrium cost of loanable funds i^e rises and hence more pools of applicants are excluded from the market. In addition, the increase in i^e raises the equilibrium interest rate $R(i^e;\mu)$ faced by all those pools of applicants with higher mean returns. Demand for loans by these pools is therefore also reduced. Thus, despite the informational asymmetry and rationing, the market adjustment proceeds in an essentially standard manner.

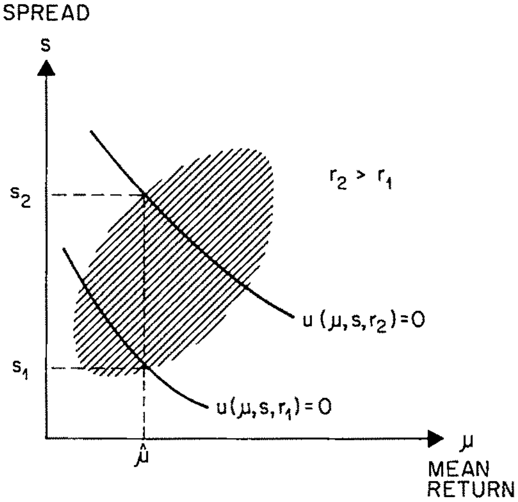


FIGURE 3. EFFECTS OF RAISING THE INTEREST RATE WITH μ AND s UNKNOWN

Could the S-W model be modified to generate more than the trivial amount of credit rationing that is implied by the above argument? Clearly, if banks find it difficult to separate out loan applicants with differing mean returns, the segmentation described here is no longer feasible. Suppose then, that the density function of a loan applicant can be written as $f(\theta,\mu,s)$, where μ is the mean return and s is a parameter reflecting the extent of the spread of the distribution around its mean. To complete the model, it is assumed that μ and s are jointly distributed according to some density function $g(\mu,s)$.

Arguing as above, the expected utility of a loan applicant facing an interest rate r, $u(\mu,s,r)$, is strictly increasing in the mean return μ and the spread s.

Figure 3 illustrates the indifference curve $u(\mu,s,r)=0$ for two different borrowing rates.

With an interest rate r_1, only those potential loan applicants on and above the curve $u(\mu, s, r_1) = 0$ will remain in the applicant pool. The remainder are better off not borrowing. In the Stiglitz-Weiss analysis, all applicants in the same risk pool have the same mean return $\hat{\mu}$. Raising the interest rate from r_1 to r_2 thus raises the minimum spread from s_1 to s_2.

Suppose instead, however, that μ and s are uniformly distributed over the shaded elliptical area in Figure 3. Then raising the interest rate not only raises the average spread, but also raises the mean return of those remaining in the applicant pool.

Loosely speaking, as long as mean and spread are *positively* correlated, the adverse effect on spread will be offset by a favorable effect on average returns. Even if mean and spread are independently distributed, the fact that the indifference curve $u(\mu, s, r) = 0$ has a negative slope implies that, under mild further assumptions, the mean return will continue to rise with the interest rate. On the other hand, if mean and spread are sufficiently negatively correlated (so that the main axis of the ellipse in Figure 3 slopes very much to the left), an increase in the interest rate lowers the mean return of those remaining in the risk pool. The adverse-selection effect analyzed by Stiglitz and Weiss is therefore reinforced. It follows that the model can indeed be modified to produce credit rationing of sufficient scope to have macroeconomic significance.

Of course, one would wish to know whether the required assumptions on the distribution of loan types were empirically plausible before drawing strong inferences from such a model. One might also ask whether, in a large pool with considerable heterogeneity, banks might find ways to screen loan applicants. Looking at the other side of the same coin, higher quality loan applicants might find ways to signal their differences by proposing contracts which differ from the simple debt contract examined here. Papers by Helmut Bester (1985) on the use of collateral and by Hellmuth Milde and myself (1987) on the use of loan size as screening devices, illustrate some of the possibilities in this regard.

REFERENCES

Bester, Helmut, "Screening vs. Rationing in Credit Markets with Imperfect Information," *American Economic Review*, September 1985, *75*, 850–55.

Milde, Hellmuth and Riley, John G., "Signalling in Credit Markets," *Quarterly Journal of Economics*, forthcoming 1987.

Stiglitz, Joseph E. and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

______ and ______, "Incentive Effects of Terminations: Applications to the Credit and Labor Markets," *American Economic Review*, December 1983, *73*, 912–27.

Credit Rationing: Reply

By JOSEPH E. STIGLITZ AND ANDREW WEISS*

John Riley has made "A Further Remark" on Section IV, "Observationally Distinguishable Borrowers," of our 1981 paper in this *Review*. In that paper, we defined two types of rationing. Criterion *a* rationing occurs when, among observationally identical borrowers, some get loans and others do not, and the rationed borrowers cannot get credit at any interest rate. A second type of rationing (criterion *b* rationing) occurs when entire types cannot get credit at any interest rate, although they would get credit if the supply of funds were sufficiently large. This type of rationing is often termed "redlining." We showed that, given the special simplifying assumptions of our model and disregarding the incentive effects of loan contracts, if there are many types of observationally distinguishable borrowers, only one type is subject to criterion *a* rationing. Riley goes on to conclude that rationing becomes of insignificant importance as the number of observationally distinct groups becomes large. More specifically, the conclusion that one might be tempted to draw from Section IV of the 1981 paper is that "Given the special simplifying assumptions of the paper, and ignoring the incentive effects of interest rates delineated in the paper, as the number of observationally distinct groups increases, the proportion of the population subject to criterion *a* rationing decreases, while the proportion subject to criterion *b* rationing decreases or increases depending on how the population is partitioned."

We did not include this conclusion (or even a less verbose version) in that paper, because we feared it might mislead readers in either of two ways. First, the reader might not have been aware that the conclusion is very sensitive to the special assumptions of the adverse-selection sections of our 1981 paper.

By the time we published our 1981 paper, our research extending that model, for example, to multiperiod settings and to situations where collateral and interest rates were both employed, had made it amply clear that the conclusion that type *a* rationing disappeared in importance as the number of types in the economy increased was not; in fact, generally valid.[1]

Second, even within the special context of those special simplifying assumptions, the conclusions would mislead a reader who was not aware that criterion *b* rationing (redlining) has consequences for allocative efficiency and macroeconomic policy that are as important as the consequences of criterion *a* rationing.

We are not so presumptuous as to believe that the assumptions we made in our 1981 paper should be interpreted as a literal description of the economy. Those assumptions were made to provide the simplest model in which the market equilibrium would be characterized by credit rationing. If one is interested (as we are) in understanding the importance of rationing in the economy, as opposed to its importance within a specific model, one should investigate the nature of rationing in more general models that are formulated to more closely resemble the actual economy.

This we have done in a series of papers, which show that with many observationally distinguishable groups there may be rationing of several, or even of *all* groups. For

*Princeton University, Princeton, NJ 08544, and Bell Communications Research, Morristown, NJ 07960, respectively.

[1] Indeed, it seemed to us transparent that even within the simpler model, the conclusion need not be valid. Assume there were some characteristic (say wealth), such that the probability distribution of returns, conditioned on that variable, did not in fact depend on that variable. Then, of course, having a finer partition of the population according to that characteristic will leave the magnitude of credit rationing unchanged.

instance, in our 1983 paper, we allowed banks and borrowers to develop multiperiod relationships. We showed that, even in the simplest dynamic model, the market equilibrium could involve rationing of both experienced and inexperienced borrowers. Rationed and nonrationed experienced borrowers were identical (differences in the results of their previous projects were due to chance), as were all inexperienced borrowers. In that dynamic model, rationing of experienced borrowers affects the choice of techniques by inexperienced borrowers. This is why both types are rationed.

In Carl Shapiro-Stiglitz (1984), unemployment of workers has similar incentive effects. In equilibrium, each type of worker has a finite probability of being unemployed. Again, this is true regardless of the number of observationally distinguishable groups in the population. Unemployment rates differ across types of workers.

In our 1985 paper, we generalized our static model to allow banks to choose simultaneously interest rates, and collateral and equity requirements. We also allowed contracts to have both sorting and incentive effects.[2] In that analysis we showed that every type of borrower could be rationed. As in the closely related work in the labor market by Stiglitz (1976), J. L. Guasch and Weiss (1980), and Barry Nalebuff and Stiglitz (1982), rationing plays a role in sorting individuals and, at the same time, is a consequence of the sorting and incentive effects of *all* the terms of the contract. The proportion of rationed borrowers does not become insignificant as the number of groups in the population increases. In our 1985 paper, pervasive rationing is possible in either a pure pooling equilibrium, in which all borrowers choose the same contracts, or in a (partial) separating equilibrium, in which the number of contracts is equal to the number of types of borrowers and there is rationing at *every* contract.

To rephrase our earlier point, we do not believe it is very interesting to explore all the ramifications of the simplifying assumptions of our 1981 paper. Rather we believe attention should be focused on understanding the robustness of the conclusions of that model when the basic assumptions are relaxed.

In investigating these questions, one must be careful to bear in mind the central economic issues. Thus, in our 1981 paper, there was only a single information problem, either adverse selection *or* moral hazard: and if there was an adverse-selection problem, individuals differed in only one respect. For this simple model, it may be possible to give the lender additional instruments, with which it can completely "solve" the informational problem without engaging in rationing. But these results provide insight only into the simplifying assumption of the model, not into the economy. In actual markets, lenders never have perfect information about the characteristics of their borrowers and can never perfectly monitor their actions. Our papers have shown that under these circumstances, credit rationing is likely to persist regardless of the number of observationally distinct groups.[3]

However, even if the specific simplifying assumptions of our 1981 paper were taken as literal descriptions of the economy, Riley's conclusions are misleading for four reasons.

First, we do not believe that the theoretical force of rationing models depends on *identical* workers being treated differently. As the number of types of borrowers increases, *nearly* identical individuals are treated differently. In the polar case of a

[2] A number of writers have constructed special examples where credit rationing does not arise. When we wrote our 1985 paper, we were well aware that an infinite number of examples can be constructed in which credit rationing does not occur. However, these examples do not vitiate our result that credit markets may be characterized by rationing. We never asserted that credit markets are always characterized by credit rationing. Much of our recent research focused on understanding the necessary and sufficient conditions for there to be rationing in credit markets and unemployment in labor markets.

[3] We are assuming that the partition of workers into more types does not eliminate problems arising from unobserved differences across borrowers and incentive effects of contracts. We believe that models in which lenders are perfectly informed concerning all the actions of each borrower are not useful for examining problems that arise due to informational asymmetries.

continuum of types of borrowers, we would find *types* of borrowers that were excluded from the credit market, though their characteristics are arbitrarily close to those of types that are getting credit. (The metric of closeness needs to be suitably defined for the economy in question.) The expected utility of types of borrowers that are excluded from the credit market is discretely lower than the expected utility of borrowers with almost identical characteristics that are not excluded.[4] Increasing the number of types of borrowers, while maintaining unobserved heterogeneity within each type, does not, in general, affect the magnitude of this problem.

Second, the proportion of borrowers who are excluded from the market—borrowers who cannot obtain loans at any interest rate even though with a larger supply of credit they would—does not necessarily change with increases in the number of types. One of the purposes of our 1981 paper was to explain this type of rationing (red-lining).

Third, we suggested in the 1981 paper that the rationing equilibria would not, in general, be Pareto efficient. The expected return of the projects of the excluded groups might exceed that of groups obtaining loans. This conclusion remains valid, even as we increase the number of types of borrowers.

Finally, Riley is incorrect in asserting that the qualitative effects of monetary policy are the same in our models as in the standard models.

One of the primary reasons for our interest in rationing equilibria is that they provide an alternative mechanism through which monetary policy may affect the level of economic activity. Though in our paper we did not have time to trace the link between actions of the monetary authority and the availability of credit (see our 1980 paper, or Alan Blinder-Stiglitz, 1983), we were concerned with how credit availability affected the level of economic activity. We stressed that it was *not* through the standard Keynesian procedure, where an increase in the supply of funds leads to a decrease in the rate of interest, which, in turn, leads to an increase in the demand for investment.

With credit rationing, an increase in the supply of funds has a direct effect, providing loans to applicants who were previously denied credit (at *any* interest rate). In particular, changes in the availability of credit affects the distribution of types of borrowers getting credit. Consequently, an outward shift in the supply of loanable funds could cause an increase in the average interest rate charged borrowers. This explains why observations on the average interest rate charged borrowers may not be helpful in determining whether monetary policy is being expansionary or contractionary.[5] These conclusions of our analysis also remain valid, regardless of the number of types of borrowers.[6]

We concluded our 1981 paper by saying:

> The Law of Supply and Demand is not in fact a law, nor should it be viewed as an assumption needed for competitive analysis. It is rather a result generated by the underlying assumptions that prices have neither sorting nor incentive effects. The usual result of economic theorizing: that prices clear markets, is model specific and is not a general property of markets—unemployment and credit rationing are not phantasms. [p. 409]

We still believe that. The objective of that paper was to construct the simplest model in which these phenomena could be explained as arising out of information imperfections

[4] By contrast, in the standard economic models, even when there are nonconvexities which result in discretely different allocations for similar types of individuals, the levels of utility are not discretely different.

[5] These problems are compounded by the fact that changes in the economic environment which lead monetary authorities to undertake expansion or contractionary actions may also lead to changes in the relationship between the expected return to the bank and the interest rate charged; leading in turn to changes in the interest rates which would be charged, at any given loan supply.

[6] Though the mechanism by which the availability of credit affects the level of investment is different from that of the standard model, it is true that an increase in the availability of credit will decrease the expected return to depositors, as in the standard model.

which we believe to be pervasive in these markets. Our subsequent research has established that the conclusion is, if anything, even more robust than we had originally thought.

REFERENCES

Blinder, Alan S. and Stiglitz, Joseph E., "Money, Credit Constraints, and Economic Activity," *American Economic Review Proceedings*, May 1983, *73*, 297–302.

Guasch, J. L. and Weiss, A., "Adverse Selection by Markets and the Advantage of Being Late," *Quarterly Journal of Economics*, May 1980, *94*, 453–66.

Nalebuff, Barry and Stiglitz, Joseph E., "Quality and Prices," Econometric Research Program Research Memorandum No. 297, May 1982.

Riley, John, "Credit Rationing, A Further Remark," *American Economics Review*, March 1987, *77*, 224–27.

Shapiro, Carl and Stiglitz, Joseph, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Stiglitz, Joseph, "Prices and Queues as Screening Devices in Competitive Markets," IMSSS Technical Report No. 212, Stanford University, August 1976.

______ **and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economics Review*, June 1981, *71*, 393–410.

______ **and** ______, "Incentive Effects of Terminations: Applications to the Credit and Labor Markets," *American Economics Review*, December 1983, *73*, 912–27.

______ **and** ______, "Credit Rationing with Collateral," Bell Communications Research Economics Discussion Paper No. 12, 1985.

______ **and** ______, "Credit Rationing in Markets with Imperfect Information, Part 1," Bell Laboratories Technical Memorandum, 1980.

ERRATUM[†]

Pechman's Tax Incidence Study: A Response

By JOSEPH A. PECHMAN*

I am indebted to Edgar Browning for calling attention to some peculiarities of the data underlying the estimates in my 1985 study. He is quite right that the ratios of transfer payments to income in the files for 1975 and later years are inconsistent with the corresponding ratios in the 1966 and 1970 files and in the *Consumer Population Surveys*, particularly in the lower part of the income distribution.

I do not believe, however, that the baby should be thrown out with the bathwater, as Browning seems to suggest. I made available to him detailed data on the tax burdens of labor income, capital income, and consumption by income classes which provided a basis for revising my calculations for 1975 and later years. Instead, he chose to draw inferences about the progressivity of U.S. taxes since 1966 on the basis of changes in the relative importance of the various taxes, which is not appropriate for this purpose. These inferences would be correct only if there were no changes in the structure of each tax and in the composition of income in the various income classes during the period studied (for example, if there were no changes in the progressivity of the individual income tax or in the distribution of dividends by income classes).[1]

It is straightforward to improve on this procedure. Where in the income distribution the share of transfer payments in total income is too low in 1975 and later years, the shares of labor and capital income are correspondingly high, and vice versa. To arrive at more realistic estimates of the shares for labor, capital, and transfer income, I first applied the 1970 shares to adjusted family income in each income decile in 1975, 1980, and 1985, and then made proportional adjustments in the columns and the rows alternately until they added to the correct totals.[2]

The incidence calculations for 1980 based on the revised weights are given in Table 1, which shows the effective rates of the major taxes by deciles.[3] These calculations are

[†]*Editors' Note*: The December 1986 publication of this response to Edgar Browning (1986) did not include Tables 1 and 2. This version is complete.

*The Brookings Institution, 1775 Massachusetts Avenue, N.W., Washington, D.C. 20036. I am grateful to Edgar Browning for helpful comments on an earlier version of this response, and to Charles E. Miles for programming the calculations.

[1] In fact, if this procedure were accurate, it would be necessary to prepare a MERGE file for only one year and then to project the incidence calculation for future years on the basis of changes in the relative importance of the various taxes used in the system.

[2] The income shares can be thought of as a matrix whose row entries add up to adjusted family income for each decile, and whose column entries add up to total capital, labor, and transfer income, respectively. As might be expected, when the 1970 shares are applied to later years, the totals do not add up to the known totals of capital, labor, and transfer income (the columns). These amounts were then adjusted proportionately to equal the known totals. This step yields incomes in each decile (the rows) which differed from the known totals, and this difference was eliminated by a proportional adjustment. These adjustments were repeated alternately until the columns (the totals of the income shares) and the rows (the total adjusted family income in each decile) added up to the known totals. For a proof that there is a unique solution to this procedure, see Michael Bacharach (1970, ch. 4). We have found that four or five iterations will yield satisfactory results, but we actually used as many as twenty to obtain the revised weights.

[3] I have not corrected the data base used in these calculations, because the costs would have been prohibitive. However, I believe that the adjustments of the weights for the various income shares described above provided a good approximation of the effective rates of the various taxes.

TABLE 1—EFFECTIVE RATES OF FEDERAL, STATE, AND LOCAL TAXES, BY POPULATION DECILE, 1980[a]

Population Decile	Individual Income Tax	Corporation Income Tax	Property Tax	Sales and Excise Taxes	Payroll Taxes	Personal Property and Motor Vehicle Taxes	Total Taxes
First[b]	2.4	1.9	2.3	8.4	2.0	0.1	17.1
Second	3.5	1.6	1.9	7.0	3.0	0.1	17.1
Third	5.1	1.3	1.6	5.9	4.7	0.1	18.9
Fourth	6.1	1.3	1.5	5.5	6.2	0.2	20.8
Fifth	7.8	1.1	1.3	5.1	7.1	0.2	22.7
Sixth	8.8	1.0	1.2	4.9	7.2	0.3	23.4
Seventh	9.9	1.2	1.3	4.6	7.1	0.3	24.4
Eighth	11.3	1.3	1.4	4.5	6.9	0.3	25.5
Ninth	12.9	1.3	1.4	3.9	6.7	0.2	26.5
Tenth	14.7	4.7	3.1	2.1	3.8	0.1	28.5
Top 5 Percent	14.8	5.9	3.7	1.6	2.7	0.1	28.9
Top 1 Percent	13.0	8.4	4.9	1.0	1.0	0.1	28.4
All Deciles[c]	10.8	2.5	2.0	4.0	5.8	0.2	25.3

Source: The Brookings MERGE file (revised).

[a] The rates are shown in percent; based on variant 1*c*, which is the most progressive set of assumptions examined in this study.

[b] Includes only units in the sixth to tenth percentiles.

[c] Includes negative incomes not shown separately.

TABLE 2—EFFECTIVE RATES OF FEDERAL, STATE, AND LOCAL TAXES, BY POPULATION DECILE, SELECTED YEARS, 1966–85[a]

Population Decile	1966	1970	1975	1980	1985
First[b]	16.8	18.8	19.7	17.1	17.0
Second	18.9	19.5	17.6	17.1	15.9
Third	21.7	20.8	18.9	18.9	18.1
Fourth	22.6	23.2	21.7	20.8	21.2
Fifth	22.8	24.0	23.5	22.7	23.4
Sixth	22.7	24.1	23.9	23.4	23.8
Seventh	22.7	24.3	24.2	24.4	24.7
Eighth	23.1	24.6	24.7	25.5	25.4
Ninth	23.3	25.0	25.4	26.5	26.2
Tenth	30.1	30.7	27.8	28.5	26.4
Top 5 Percent	32.7	33.0	28.4	28.9	26.0
Top 1 Percent	39.6	39.0	29.0	28.4	25.3
All Deciles[c]	25.2	26.1	25.0	25.3	24.5

Source: The Brookings MERGE files (revised).

[a] The rates are shown in percent; based on variant 1*c*, the most progressive set of incidence assumptions used in this study.

[b,c] See Table 1.

based on the most progressive set of assumptions (variant 1*c*) presented in my study. Variant 1*c* assumes that the corporation income and property taxes are borne by capital in general, the payroll tax by labor, consumption taxes by consumers, and the individual income tax by those who pay it. I believe that this variant more nearly reflects the present state of incidence theory than any of the other variants.[4]

[4] Browning agrees, but he would go further. He would allocate consumption taxes to factor income rather than

The revised calculations for 1980 give much more plausible estimates not only of the distribution of payroll taxes, but of other taxes as well. As in previous years, the effective payroll tax rate rises in the lower half of the income distribution, because transfer payments decline as a percentage of income; regressivity sets in thereafter. The U-shaped incidence by income classes of the corporation income and property taxes is also restored.[5]

Changes in the overall progressivity of the tax system for the period covered by the MERGE files are shown in Table 2 on the basis of the revised estimates. The distinct reduction from 1966 to 1985 in the effective tax rates in the top part of the distribution shown in the original calculations remains in the revised estimates. In 1966, those in the top decile paid 30.1 percent of their incomes in taxes; by 1985, their tax burden had declined to 26.4 percent. In the top percentile, the decline was from 39.6 percent to 25.3 percent. This was the result largely of the reductions in the corporation income tax and the decline in the importance of the property tax during this period.

On the other hand, the increase in tax burdens in the bottom part of the distribution shown in the original data disappears in the revised data. Although the effective tax rate in the bottom decile increases between 1966 and 1975, it declines between 1975 and 1985. On balance, the large increase in the payroll tax from 1966 to 1985 is just about offset by the effect of the (revised) increase in transfer payments in the first decile and more than offset in the second. However, if families were classified by market incomes rather than income plus transfer payments, the increase in the tax burdens at the lower end of the distribution resulting from the higher payroll taxes would be pronounced.[6]

Basically, the revised estimates show that there has been very little change in tax burdens throughout most of the income distribution, except at the very top where tax burdens have declined sharply because of reduced taxes on capital.[7]

A set of the revisions for the key tables for 1975, 1980, and 1985 is available on request.

to consumption (see his paper, 1985), but he has not yet persuaded the profession of his position.

[5] The relatively high income tax ratios in the lower deciles are the results of the accrual concept of income used in the analysis. Employee compensation includes employer contributions to pension plans in the year contributed, but the tax is deferred until pension benefits are received. Since there is no income to correspond to the employer contribution in the year the pension is received, the effective tax rate of persons receiving pensions is relatively high. For a discussion of this and other reasons why effective rates based on an annual accounting period may be distorted, see my study (pp. 48–50).

[6] The combined effect of transfer payments and taxation on the distribution of market incomes is shown in my study (table 4–7 and figure 4). On the basis of this classification, the tax transfer system is highly progressive, because transfers add to incomes in the bottom part of the distribution while taxes subtract from them in the top part. Unfortunately, we did not tabulate the 1966 and 1970 data by market incomes and therefore cannot make the comparison over the longer period 1966–85.

[7] Calculations using 1966 instead of 1970 income shares as weights raise progressivity for 1975, 1980, 1985, but the change from 1965–85 remains relatively small except at the very top of distribution. On the basis of the 1966 weights, the 1985 effective tax rates increase from 14.5 percent in the bottom decile to 24.1 percent in the sixth decile, and 26.0 percent in the ninth decile, and then decline to 25.6 percent in the top decile, and 25.2 percent in the top percentile.

REFERENCES

Bacharach, Michael, *Biproportional Matrices and Input-Output Change,* Cambridge: Cambridge University Press, 1970.

Browning, Edgar K., "Tax Incidence, Indirect Taxes and Transfers," *National Tax Journal,* December 1985, *38,* 525–534.

______, "Pechman's Tax Incidence Study: A Note on the Data," *American Economic Review,* December 1986, *76,* 1214–18.

Pechman, Joseph A., *Who Paid the Taxes, 1966–85?,* Washington: The Brookings Institution, 1985.

NOTES

The American Economic Association announces the *Journal of Economic Perspectives*: Joseph E. Stiglitz, Editor; Carl Shapiro, Co-Editor. The *JEP* is a new quarterly journal to be published by the AEA. Members will automatically receive the first issue scheduled to appear mid-1987. The journal's mission is to provide economists with accessible articles that report on and critique recent research findings, evaluate public policy initiatives, and serve as insightful readings for classroom use. The editors intend that the *JEP* will facilitate the diffusion of current research not only within the academic sphere, but also throughout the public sector and the business community. All articles will be commissioned by the Editorial Board. Members of the Editorial Board are Henry J. Aaron, Stanley Fischer, Paul R. Krugman, Edward P. Lazear, Mark J. Machina, Charles F. Manski, Donald N. McCloskey, Bernard Saffran, Steven C. Salop, Lawrence H. Summers, Hal R. Varian, and Janet L. Yellen. The editorial office address is *JEP*, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544.

After the death of Daniel Saks in January 1986, his family, friends, colleagues, and former students established a fund to honor his memory. It is hoped that sufficient funds will be collected to endow the fund so that an annual lecture, directed to students and faculty, can be delivered by a distinguished scholar on issues in educational research and policy. These lectures will be published and distributed widely to continue to promote research on topics that were important to Saks' professional interests. The first lecture will be delivered March 16, 1987, by Professor Lee Schulman of Stanford University on "Metaphors From the Disciplines: Multiple Perspectives on Teaching and Policy." Contributions to the fund are tax deductible. They should be marked "Daniel Saks Memorial Fund, George Peabody College," and sent to Dean Willis Hawley, Box 329, Peabody College, Vanderbilt University, Nashville, TN 37203.

New Address: The editorial offices of the *Journal of Macroeconomics*: College of Business Administration, Louisiana State University, Baton Rouge, LA 70803.

The Duke University Manuscript Department has established an Economists' Papers Project to acquire and preserve the correspondence, writings, and other papers of a select number of especially distinguished economists. Among the economists who have committed their papers to Duke are Kenneth J. Arrow, Lawrence R. Klein, H. Gregg Lewis, Robert E. Lucas, and Lionel W. McKenzie. Advisors to the project are professors A. W. Coats, Neil de Marchi, Craufurd Goodwin, and E. Roy Weintraub. Persons interested in the project or aware of papers that might be suitable for preservation at Duke should contact one of the advisors, or Robert L. Byrd, Curator of Manuscripts, Duke University Library, Durham, NC 27706 (telephone 919+684-3372).

The United States Institute of Peace announces interim procedures for grant applications to implement research, education, and training "to promote international peace." The Institute will not fund grant proposals of a partisan political nature, or those that would bring the Institute into the policymaking processes of any government or government agency. The grants may be issued to individuals, nonprofit organizations, and official public institutions. For full information and application forms, contact United States Institute of Peace, 730 Jackson Place, NW, Washington, D.C. 20503 (telephone 202+789-5700).

The Council for International Exchange of Scholars (CIES) announces that a number of 1987-88 Fulbright Lecturing Grants remain available to U.S. faculty in the field of economics. There are specific openings in Botswana, Bulgaria, Burundi, Central African Republic, People's Republic of China, Colombia, Czechoslovakia, Egypt, Hungary, Indonesia, Lesotho, Malaysia, Mexico, Morocco, Niger, Nigeria, Papua New Guinea, Peru, Philippines, Poland, Portugal, Romania, Senegal, Sudan, Taiwan, Tanzania, Turkey, USSR, Venezuela, West Bank, and Yemen. In addition, other countries are open to applications in any discipline, and economics is among their preferred fields. Scholars in all academic ranks, including emeritus, are eligible to apply and it is expected that applicants will have a Ph.D., college, or university teaching experience, and evidence of scholarly productivity; U.S. citizenship is required. In a few countries (of Central and South America and Francophone Africa), knowledge of the host country language is required. For information, call or write CIES, Eleven Dupont Circle NW, Suite 300, Washington, D.C. 20036 (telephone 202+939-5401). When inquiring, indicate countries of interest.

The Credit Research Center at the Krannert School of Management, Purdue University, has funded an annual award of $1,500 for research in the area of consumer credit. To be eligible, submit a completed research paper on a consumer credit or markets topic postmarked by April 1, 1987. The winner will be determined by a panel made up of the staff of the Credit

Research Center and members of the Advisory Council of the Credit Research Center. By submitting a paper for this competition, the author(s) certify that the paper has not won similar awards or been published before. Send papers to A. Charlene Sullivan, Acting Director, Credit Research Center, Purdue University, West Lafayette, IN 47907 (telephone 317+494–4380).

The S. S. Huebner Foundation is seeking applicants for pre- and postdoctoral fellowships. Fellowships provide full tuition and a stipend of $800 per month for study of risk and insurance at the Wharton School, University of Pennsylvania. Postdoctoral fellowships are available for one year, while the predoctoral fellowships are renewable up to four years. Further information can be obtained from Professor J. David Cummins, Wharton School, 3641 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104–6218.

The Graduate School of Business, University of Florida, announces the establishment of the James W. Walter Eminent Scholar Chair in Entrepreneurship. The Search Committee invites nominations and applications of persons with outstanding competence from diverse research and teaching backgrounds, including Economics, Industrial Organization, Public Policy, Business Policy, Financial Management, or other areas that study the growth, development, and management of business and other economic entities. March 15, 1987, is the closing date for applications and nominations to be sent to Robert F. Lanzillotti, Chair, Search Committee, James Walter Chair of Entrepreneurship, Graduate School of Business, University of Florida, Gainesville, FL 32611.

The National Center for Food and Agricultural Policy at Resources for the Future (RFF) is offering resident fellowships to individuals holding the Ph.D. in any discipline, employed by university, government, or the private sector. Professionals who will be on sabbatical or other leave of absence are encouraged to apply. Stipend negotiated individually, plus research support, office facilities at RFF, and allowance for moving expenses. Fellowships awarded for six to twelve months. The deadline is May 1, 1987.

The Albert Gallatin Fellowship International Affairs provides for nine months of study at the Graduate Institute of International Studies, University of Geneva, Switzerland, by an American candidate for the Ph.D. who is actively engaged in dissertation research for the doctorate. The Fellowship provides a stipend of $7,470 for living and other expenses for the academic year October 1987–July 1988; roundtrip travel New York-Geneva; an allowance for travel outside of Geneva, if required by fellow's research, to be determined in consultation with the fellow's supervisor; an allowance for purchase or transport of books related to research. Application forms and full information may be obtained from William D. Carter, FERIS Foundation of America, 5 Harvard Court, Huntington, NY 11743 (telephone 516+271–1762). The closing date for receipt of applications is March 10, 1987.

The half-centenary Congress of the International Institute of Public Finance (IIPF) will be held in Paris, August 24–28, 1987, dealing with Public Finances and Performance of Enterprises. Inquiries should be addressed to the chairman of the scientific committee, Professor Manfred Neumann, University Erlangen-Nurnberg, P.O.B. 3931, D–8500 Nurnberg, West Germany, or to the President of IIPF, Professor Karl W. Roskamp, Wayne State University, Detroit, MI 48202.

The annual Conference of the Association of Private Enterprise Education will be held in Atlanta, Georgia, April 26–29, 1987. The emphasis of the meeting is on private enterprise and entrepreneurship. For full information, contact Dr. Calvin Kent, Hankamer School of Business, Baylor University, Waco, TX 76798.

The Language and International Trade Program, with the cosponsorship of Eastern Michigan University's College of Business and the Arts & Humanities Service Center, announces the sixth annual EMU Conference on Languages and Communication for World Business and the Professions, May 8–9, 1987, Ann Arbor, Michigan. For full information, contact Dr. John R. Hubbard, Conference Co-Chairman or Dr. David Victor, Assistant Professor of Management, c/o Language and International Trade Program, 217 Alexander Building, Eastern Michigan University, Ypsilanti, MI 48197 (telephone 313+487–0178/–2283/–0130).

The seventh Berkshire Conference on the History of Women will be held at Wellesley College, Wellesley, Massachusetts, June 19–21, 1987. For registration and further information, contact Ms. Jean Proctor, Berkshire Conference, Women's Studies Program, Wellesley College, Wellesley, MA 02181 (telephone 617+235–0320).

The first International Conference on Statistical Data Analysis Based on the L_1 Norm and Related Methods will be held August 31–September 4, 1987, at the University of Neuchatel, Switzerland. For further information, registration materials, and program details, please contact Professor Yadolah Dodge, Conference Organizer, Universite de Neuchatel, Groupe d'Informatique et de Statistique, Pierre-a-Mazel 7, CH-2000 Neuchatel, Switzerland (telephone 038+25 72 05).

A NATO Advanced Study Institute on Games with Incomplete Information and Bounded Rationality Decision Models will be held at the Villa Monastero on Lake Como in Varenna, Italy, June 4-14, 1987. For further information, contact H. W. Kuhn, Department of Mathematics, Fine Hall-Washington Road, Princeton University, Princeton, NJ 08544, or G. Gambarelli, University of Bergamo, Via Salvecchio, 19; Bergamo 24100, Italy.

The Korea Development Institute offers fellowships to American professors and Ph.D. candidates conducting research related to the Korean economy. Awards cover six months to one year of residence in Seoul. The Ph.D. candidates will normally receive $15,000 plus one roundtrip air ticket. Faculty members will normally receive $2,000 per month plus two roundtrip air fares and an apartment. Fellows will also be given office space and access to KDI's excellent library and computer facilities. Knowledge of the Korean language is not necessary. The Ph.D. applications are normally accepted in February and August while faculty members may apply at any time. Further information can be obtained from Dr. Jwa Sung Heo, Korea Development Institute, Box 113, Cheongryong, Seoul, Korea, or from Professor Leroy P. Jones, Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215.

The Council for International Exchange of Scholars (CIES) announces the competition for the 1988-89 Fulbright grants in research and university lecturing abroad. The awards include more than 300 grants in research and 700 grants in university lecturing for periods ranging from three months to a full academic year. There are openings in over 100 countries and, in some instances, the opportunity for multi-country research is available. Fulbright awards are granted in virtually all disciplines, and scholars in all academic ranks are eligible to apply. Applications are also encouraged from retired faculty and independent scholars. Benefits include roundtrip travel for the grantee and, for most full academic year awards, one dependent; maintenance allowance to cover living costs of grantee and family; tuition allowance, in many countries, for school-age children; and book and baggage allowances.

The basic eligibility requirements for a Fulbright Award are U.S. citizenship; Ph.D. or comparable professional qualifications; university or college teaching experience; and, for selected assignments, proficiency in a foreign language. Note that a new policy removes the limit of two Fulbright grants to a single scholar. Application deadlines are: June 15, 1987 (for Australia, India, and Latin America, except lecturing awards to Mexico, Venezuela, and the Caribbean); September 15, 1987 (for Africa, Asia, Europe, the Middle East, and lecturing awards to Mexico, Venezuela, and the Caribbean); November 1, 1987 (for institutional proposals for the Scholar-in-Residence Program); January 1, 1988 (for Administrators' Awards in Germany, Japan, and the United Kingdom; the Seminar in German Civilization; the NATO Research Fellowships, and the Spain Research Fellowships); and February 1, 1988 (for the France, Italy, and Germany Travel-Only Awards). For more information and applications, call or write CIES, Eleven Dupont Circle N.W., Washington, D.C. 20036-1257 (telephone 202+939-5401).

The CIES has two directories of interest to readers: The *Directory of American Scholars 1986-87* lists all American current Fulbright grantees by discipline, including a two-page section which groups all the economists who have received the awards, as well as their host countries abroad; and the *Directory of Visiting Fulbright Scholars 1986-87* lists all foreign Fulbright Scholars currently in the United States. Included is a section listing about 60 economists from abroad and their host institutions in this country. For full information, contact CIES (telephone 202+939-5401).

The Indo-U.S. Subcommission on Education and Culture is offering twelve long-term (6-10 months) and nine short-term (2-3 months) awards for 1988-89 research in India. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. The fellowship program seeks to open new channels of communication between academic and professional groups in the United States and India. Therefore, scholars and professionals with limited or no prior experience in India are especially encouraged to apply. Fellowship terms include: $1,500 per month, of which $350 per month is payable in dollars and the balance in rupees; an allowance for books and study/travel in India; and international travel for the grantee. In addition, long-term fellows receive international travel for dependents; a dependent allowance of $100-250 per month in rupees; and a supplementary research allowance up to 34,000 rupees. The application deadline is June 15, 1987. For application forms and further information contact Council for International Exchange of Scholars, Att: Indo-American Fellowship Program, Eleven Dupont Circle, Suite 300, Washington, D.C. 20036-1257 (telephone 202+939-5469).

A Conference on the Stockholm School After 50 Years will be held in Stockholm, August 31-September 1, 1987. The purpose is to examine the contributions of the Stockholm School. Those wishing to participate or contribute should write to Lars Jonung or Klas Fregert, Department of Economics, University of Lund, Box 5137, S-220 05 Lund, Sweden.

Call for Papers: The editors of the *Kentucky Journal of Economics and Business* solicit papers for their 1987 annual edition. Topics may cover any area related to economics, and are not limited to a particular region of the country. To submit a paper or obtain further information, contact Professor John Bethune, Bellarmine College, Louisville, KY 40205 (telephone 502+452–8240).

Call for Papers: The *Journal of Human Resources* is planning a special issue on New Methods in Economic Demography. Papers are sought that apply frontier econometric techniques to issues in economic demography. Some topic areas are fertility, marital dissolution and formation, household composition in general, migration, and the role of labor supply in each of these areas. The deadline is August 31, 1987. Please submit manuscripts to the *Journal of Human Resources*, Social Sciences Building, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

Call for Papers: A special issue of the *Journal of Aging Studies* will be published in the winter of 1988. The topic is "Interpretations of Social Security," widely defined to include not only analyses of the program and its development but also studies examining the impact of all Social Security programs—OASDI, Medicare, Medicaid and SSI—on individuals, families, communities, and social services. Comparative studies are also welcome. The deadline is September 20, 1987 (maximum 35 pages, including bibliography, tables or figures). Send two copies to the guest editor, Professor Jill Quadagno, Department of Sociology, University of Kansas, Lawrence, KS 66045.

Call for Papers: The European Association for Research in Industrial Economics (EARIE) will hold its annual conference in Madrid, Spain, August 30–September 1, 1987. To present a paper, send a one-page abstract and three copies of the paper by April 1 to Oscar Fanjul, EARIE Programme Chairman, Alberto Alcocer 47, 28016 Madrid, Spain.

Call for Papers: The European Finance Association (EFA) will hold its annual meeting in Madrid, Spain, September 3–5, 1987. To present a paper, send it or an abstract by April 1 to Professor Angel Berges, Facultad Economicas, Universidad Autonoma, Madrid, Spain 28049. (Use the same address to serve as chair or discussant.)

Call for Papers: The first annual meeting of the European Society for Population Economics (ESPE) will be held at the Erasmus Universiteit Rotterdam Campus, in Rotterdam, September 18–19, 1987. Those interested in participating should submit their papers to Pierre Pestieau, ESPE Program Chairman, Department of Public Economics, Universite de Liege–Sart-Tilman, 7, Boulevard du Rectorat (B.31), 4000–Liege, Belgium. The deadline for submission is April 1, 1987.

Call for Papers: The Australian Conference of Health Economists will be held at the Australian National University, Canberra, September 24–25, 1987. The conference format requires that printed papers be distributed to participants prior to the conference and discussants are appointed for each paper accepted for presentation. Papers on any topic applying economic theory to issues in the health sector are welcome. Further information can be obtained from the editors of the conference *Proceedings*, D. P. Doessel, Department of Economics, University of Queensland, St. Lucia, Queensland 4067, Australia, or J.R.G. Butler, Department of Marketing and Applied Economics, Brisbane C.A.E., P.O. Box 117, Kedron, Queensland 4031, Australia.

Call for Papers: The Association for the Social Sciences in Health (ASSH) seeks papers for its sessions at the annual meeting of the American Public Health Association in New Orleans, Louisiana, October 18–22, 1987. Abstracts must be submitted by March 27, 1987, on the standard abstract form, which appears in the January 1987 *American Journal of Public Health*, or the APHA newsletter, *The Nation's Health*. Please send six copies, one camera ready and five photocopies, and a self-addressed, stamped envelope to Linda A. Siegenthaler, National Center for Health Services Research, Room 18A–19, #9, 5600 Fishers Lane, Rockville, MD 20857.

Call for Papers: Risk-Based Payments for Health Care under Public Programs: Toward a Core of Knowledge for Expanded Choices, a conference supported by the U.S. Health Care Financing Administration, will be held in Williamsburg, Virginia, October 8–9, 1987. The purpose is to highlight research on alternative financing and delivery systems for Medicare, Medicaid, and other public programs. The deadline for abstracts is April 1, 1987. For further information (or to be a discussant), contact Louis F. Rossiter, Associate Professor of Health Economics, Department of Health Administration, Medical College of Virginia, Virginia Commonwealth University, P.O. Box 203, Richmond, VA 23298–0001.

Call for Papers: The annual meeting of the Association of Managerial Economists (AME) will be held in Chicago, Illinois, December 28–30, 1987. Three sessions of competitively selected papers will be featured in conjunction with the ASSA meetings. Theoretical, empirical, and policy analyses across a broad spectrum of topics will be featured, including theoretical and empirical studies of advertising, competitive strategy, diversification, financial decisions, forecasting, innovation, and pricing. Especially encouraged are papers integrating the theory of accounting, finance and industrial organization. Both members and nonmembers are invited to submit papers and/or make program suggestions to Professor Mark Hirschey, AME Program Chair, Jones Graduate School of Administration, Rice University, P.O. Box 1892, Houston, TX 77251 (telephone 713+527-8101, ext. 2549). Submission deadline is July 15, 1987.

Call for Papers: The annual meeting of the Association of Environmental and Resource Economists (AERE) will be held jointly with the AEA in Chicago, December 28–30, 1987. Those interested in having papers considered for these sessions should send three copies of a one-page abstract to Robert Halvorsen, Chair of AERE's Contributed Abstract Committee, Department of Economics, University of Washington, Seattle, WA 98195, no later than June 15, 1987.

The National Institute on Aging and the National Institute of Mental Health Task Force on Longitudinal Research Methods announces the availability of five annotated, brief bibliographies of longitudinal research methodologies. The bibliographies describe some of the core publications in selected areas of longitudinal research methods: *Longitudinal Factor Analysis* (John Tisak and William Meredith); *Event History Analysis* (Jan M. Hoem); *Longitudinal Structural Equation Modeling* (J. J. McArdle); *Quantitative Measures* (David Rogosa); *Single Case Designs and Data Analysis* (John Nesselroade). To obtain copies of the set, please send a mailing label with your name and address to Dr. Ronald P. Abeles, Bibliographies, Behavioral Sciences Research, National Institute on Aging, Building 31, Room 4C32, Bethesda, MD 20892.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45th Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Anna Koutsoyiannis, professor of economics, University of Ottawa, September 10, 1986.

Foreign Scholars

Michael Matsebula, University of Swaziland: visiting foreign scholar, department of economics, University of Delaware, September 1, 1986–August 31, 1987.

Promotions

James N. Dertouzos: associate head, economics and statistics department, Rand Corporation, November 1985.

Paul Dileo: economist, external financing department, Developing Economies Division, Federal Reserve Bank of New York, July 24, 1986.

Gary J. Dorman: vice president, National Economic Research Associates, November 1, 1986.

Barry L. Falk: associate professor of economics, Iowa State University, August 1986.

Phillip Fanchon: associate professor of economics, University of Nevada-Reno, July 1986.

John Geanakoplos: professor of economics, Yale University, July 1, 1986.

Roger Ginger: professor of economics, Iowa State University, July 1, 1986.

Kauser Hamdani: senior economist, domestic research department, Domestic Research Division, Federal Reserve Bank of New York, August 21, 1986.

Hugo Mario Hervitz: associate professor of economics, Barry University-Miami, September 1986.

James R. Hosek: head, department of economics and statistics, Rand Corporation, November 1985.

Sumner J. LaCroix: associate professor of economics, University of Hawaii-Manoa.

Robert T. McGee: senior economist, financial studies department, Domestic Financial Markets Division, Federal Reserve Bank of New York, August 21, 1986.

Louis Maccini: professor of economics, Johns Hopkins University, July 1, 1986.

Warren Moskowitz: senior economist, international affairs department, country risk studies staff, Federal Reserve Bank of New York, September 18, 1986.

Daniel M. Otto: associate professor of economics, Iowa State University, July 1, 1986.

Marcellus S. Snow: professor of economics, University of Hawaii-Manoa, July 1986.

Charles Steindel: senior economist, financial studies department, Domestic Financial Markets Division, Federal Reserve Bank of New York, August 21, 1986.

Gary W. Williams: associate professor of economics, Iowa State University, July 1, 1986.

Administrative Appointments

Robert Archibald: chair, department of economics, College of William and Mary, July 1, 1986.

John R. Finlay: acting dean, School of Business and Economics, Wilfrid Laurier University, January 1–June 30, 1987.

James S. Hanson: chairman, department of economics, Willamette University, July 1985.

Joseph A. McKenzie: director, Policy Analysis Division, Office of Policy and Economic Research, Federal Home Loan Bank Board, April 1986.

J. Peter Matilla: acting director, Industrial Relations Center, Iowa State University, August 1986–June 1987.

Frank W. Millerd: chairman, department of economics, Wilfrid Laurier University, January 1, 1987.

William D. Nordhaus: provost, Yale University, July 1, 1986.

Rulon D. Pope: chairman, department of economics, Brigham Young University, fall 1986.

New Appointments

Faye Anderson, Purdue University: assistant professor, department of economics, Colorado State University, August 20, 1986.

Beth J. Asch, Center for Naval Analyses: associate economist, Rand Corporation, March 1986.

Stacie Beck: assistant professor, department of economics, University of Delaware, September 1, 1986.

Timothy Brennan, Antitrust Division, U.S. Department of Justice: associate professor, telecommunications policy and economics, George Washington University, September 1986.

Richard C. K. Burdekin: assistant professor, department of economics, University of Miami, August 1986.

Harold C. Cochrane: associate professor, department of economics, Colorado State University, July 1, 1986.

Dennis DiPietre: temporary assistant professor, department of economics, Iowa State University, 1986–87.

Phillip Dybvig: professor of economics and organization, Yale University, July 1, 1986.

Evangelos Falaris, Ohio State University: assistant professor, department of economics, University of Delaware, September 1, 1986.

Danial J. Feaster: assistant professor, department of economics, University of Miami, August 1986.

Klaus K. Frohberg: adjunct associate professor, department of economics, Iowa State University, October 1, 1986.

B. Delworth Gardner, University of California-Davis: department of economics, Brigham Young University, fall 1986.

Vittorio Grilli, University of Rochester: assistant professor, department of economics, Yale University, July 1, 1986.

Adam J. Grossberg: assistant professor, department of economics, Trinity College, September 1, 1986.

John Haltiwanger: associate professor, department of economics, Johns Hopkins University, July 1, 1986.

Koichi Hamada, University of Tokyo: professor of economics, Yale University, July 1, 1986.

Edward A. Hjerpe III, Commodity Futures Trading Commission: financial economist, Office of Policy and Economic Research, Federal Home Loan Bank Board, June 1986.

Beverly Hirtle: economist, financial studies department, Domestic Financial Markets Division, Federal Reserve Bank of New York, September 8, 1986.

Todd L. Idson: assistant professor, department of economics, University of Miami, August 1986.

Stephen Karlson, Wayne State University: associate professor, department of economics, Northern Illinois University, August 1986.

Stephen King: economist, financial studies department, special financial studies staff, Federal Reserve Bank of New York, September 22, 1986.

James B. Kliebenstein: associate professor, department of economics, Iowa State University, August 1, 1986.

Bruce H. Kobayashi, University of California-Los Angeles: economist, economic litigation section, Antitrust Division, U.S. Department of Justice, September 1986.

Daniel Y. Lee, University of Pittsburgh: associate professor, department of economics, Shippensburg University, August 28, 1986.

Nancy Lutz, Stanford University: assistant professor, Yale University, July 1, 1986.

James McGlone, Virginia Polytechnic Institute and State University: assistant professor, Northern Illinois University, August 1985.

Rosa Matzkin, University of Minnesota: assistant professor, Yale University, July 1, 1986.

Michael P. Murray: Rand Corporation and Claremont Graduate School: professor of economics, Bates College, September 1986.

Edgar Norton, Northwest Missouri State University: associate professor, department of economics, Liberty University, August 1, 1986.

Kent H. Osband, University of California-Berkeley: associate economist, Rand Corporation, September 1986.

Thomas Paynter, Northern Illinois University: Illinois State Commerce Commission, May 15, 1986.

Steven W. Popper, University of California-Berkeley: associate economist, Rand Corporation, September 1986.

Subroto Roy, Brigham Young University: assistant professor, department of economics, University of Hawaii-Manoa, August 1986.

Paul A. Samuelson: visiting professor of political economy, Center for Japan–U.S. Business and Economic Studies, New York University, October–December, 1986 and February–May, 1987.

Charles Schorin: economist, international finance department, International Financial Markets Division, Federal Reserve Bank of New York, September 8, 1986.

William Sjostrom, University of Washington: assistant professor, Northern Illinois University, August 1986.

Byung-Hee Soh, Oklahoma State University: visiting assistant professor, Northern Illinois University, August 1986.

David J. Smyth, Wayne State University: professor of economics, Louisiana State University, January 1, 1987.

David E. Spencer, Washington State University: department of economics, Brigham Young University, fall 1986.

Duncan Thomas, Princeton University: assistant professor, Yale University, July 1, 1986.

David Torregrosa: visiting assistant professor, department of economics, College of William and Mary, August 1986–May 1987.

Robert Triest: instructor, Johns Hopkins University, July 1, 1986.

Sally Jane Van Ciclen, Princeton University: economist, economic litigation section, Antitrust Division, U.S. Department of Justice, October 1986.

Cathleen L. Whiting, University of Washington: assistant professor of economics, Willamette University, September 1986.

Leaves for Special Appointments

John H. Crockett, George Mason University: visiting scholar, Office of Policy and Economic Research, Federal Home Loan Bank Board, September 1986–August 1987.

A. Frank Thompson, University of Cincinnati: visiting scholar, Office of Policy and Economic Research, Federal Home Loan Bank Board, September 1986–August 1987.

D. Ward Mardfin, Hawaii Loa College: Fulbright professor of economics, University of the South Pacific, December 1987.

Lawrence J. White, New York University: Board Member, Federal Home Loan Bank Board, November 12, 1986.

Resignations

John J. Bigelow, Yale University: University of Iowa, June 30, 1986.

Gregory Dow, Yale University: University of Alberta, June 30, 1986.

Tatsuro Ichiishi, University of Iowa, January 8, 1987.

Tatsuo Hatta, Johns Hopkins University, December 30, 1985.

Kathy J. Hayes, Northern Illinois University: Southern Methodist University, August 1986.

Masahiro Kawai, Johns Hopkins University, June 30, 1986.

Richard R. Nelson, Yale University: Columbia University, June 30, 1986.

Miscellaneous

Mohan Munasinghe, Senior Energy Advisor to the President of Sri Lanka, 1986 International Merit Award, International Association of Energy Economics.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Editorial Office, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542-4607.

NOTICE TO ALL GRADUATE DEPARTMENTS

PLEASE NOTE: The *American Economic Review* will no longer carry the annual list of doctoral dissertation recipients published by the AEA.

The *Journal of Economic Literature* will print the eighty-fourth list of recipients and titles of degrees conferred in economics in U.S. and Canadian universities and colleges for the academic year ending June 1987. This announcement is an invitation to send them information for the preparation of the list.

By June 30, 1987, please send individual 3×5 cards, conforming to the style shown below. Please use the *JEL* classification system that is printed in each issue of the *JEL.*

ALL ITEMS AND INFORMATION FOR THE 1987 LIST SHOULD BE SENT TO:

The Journal of Economic Literature, P.O. Box 7320, Oakland Station, Pittsburgh, PA 15213.

JEL Classification No. ________

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ________________________________

Degree Conferred (Ph.D. or D.B.A.) ____________ Year ____________________

Dissertation Title: ____

Computer Access to Articles in the JEL Subject Index

Online computer access to the *JEL* and *Index of Economic Articles* database of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 *(Economic Literature Index)* contains complete bibliographic citations to articles from the nearly 300 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts published in *JEL* since June 1984 are also available as part of the full bibliographic record. The *Economic Literature Index* also includes citations to articles in the 1979 and 1980 collective volumes (collected papers, proceedings, etc.) for the *Index* database; other years will be added as soon as completed. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers. (For a complete description of the *Economic Literature Index* with search examples and suggestions for searching techniques, see the article "Online Information Retrieval for Economists–The Economic Literature Index," in the December 1985 issue of the *Journal of Economic Literature.)*

Access Options:

- **DIALOG** offers a variety of contract choices, including the option (for a low annual fee) to pay for only what you use. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call: DIALOG Information Services, Inc. Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800-3-DIALOG or 800-334-2564).

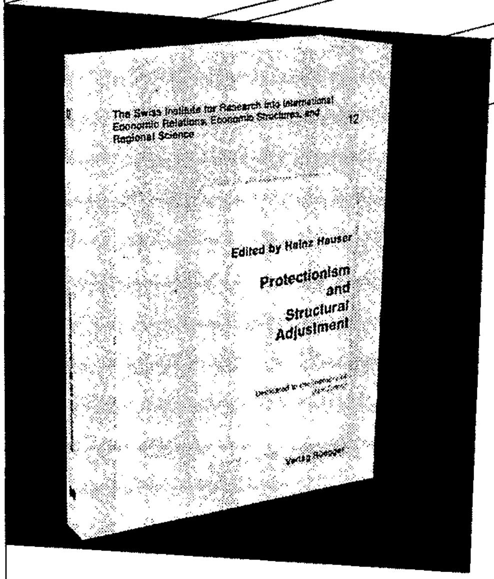
- **EasyNet,** a gateway service, provides menus to guide the untrained user through database searches in DIALOG and other databases. For information, call 1-800-841-9553 or dial up **EasyNet** on your terminal (1-800-EASYNET) and pay for your search by credit card.

Classroom Instruction:

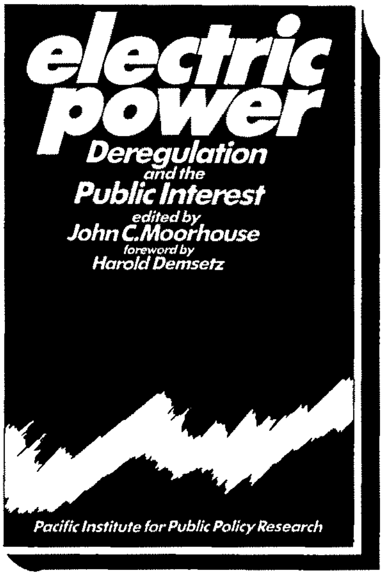
- DIALOG's Classroom Instruction Program, available at a special rate of $15/connect hour to academic institutions for supervised instruction, permits teachers to incorporate online bibliographic searching in their courses. For information, contact DIALOG or your librarian.

Shared Access:

- Economics departments or other groups may wish to consider low-cost communications software products that offer internal accounting to track the use of a single password by multiple users. Such software may permit groups to take advantage of quantity discounts or other favorable subscription arrangements.


The Swiss Institute for Research into International Economic Relations, Economic Structures, and Regional Science
12
Edited by Heinz Hauser
Protectionism and Structural Adjustment
Verlag Rüegger

DEREGULATION AND THE PUBLIC INTEREST


electric power
Deregulation and the Public Interest
edited by John C. Moorhouse
foreword by Harold Demsetz
Pacific Institute for Public Policy Research

ELECTRIC POWER

Deregulation and the Public Interest

Edited by JOHN C. MOORHOUSE

Foreword by HAROLD DEMSETZ

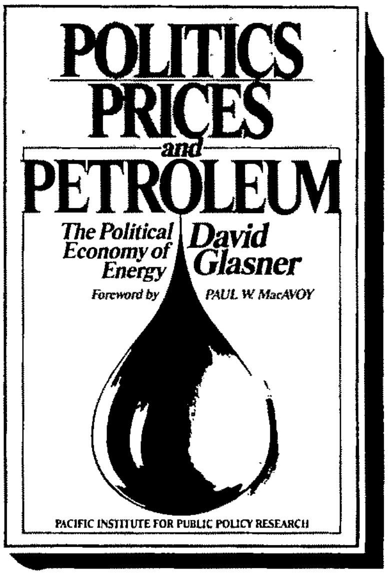
Sixteen energy economists examine the inflexible and outmoded regulatory structure of the electric power industry and challenge the scale economies underlying the traditional, static "natural monopoly" theory of utilities. John Moorhouse and his colleagues analyze utility commission behavior, investment decisions, strategies for environmental protection, and the potential for deregulation of the electric power industry.

Electric Power demonstrates that governmental regulatory and subsidy policies have led directly to the many real-world problems facing the industry today. Deregulating the electricity markets, say the authors, will improve interfuel competition, and create opportunities for real solutions–decentralized self-generation, co-generation, direct competition (for heavy industrial users) between electric utilities, and bulk wheeling.

"An excellent study that combines empirical work with modern economic analysis."
–KENNETH G. ELZINGA
University of Virginia

"An extraordinarily stimulating collection of essays on some extremely controversial topics."
–JOHN ANDERSON
Electricity Consumers Resource Council

500 pages, cloth 0-936488-11-5, $34.95
paper, 0-936488-02-6, $14.95


POLITICS PRICES and PETROLEUM
The Political Economy of Energy
David Glasner
Foreword by PAUL W. MacAVOY
PACIFIC INSTITUTE FOR PUBLIC POLICY RESEARCH

POLITICS, PRICES, AND PETROLEUM

The Political Economy of Energy

By DAVID GLASNER

Foreword by PAUL W. MACAVOY

Introduction by BENJAMIN ZYCHER

This volume demonstrates that the years of chaos in the petroleum market (1973–1979) were the inevitable consequence of a chain of faulty government actions that began in the 1950s, grew to a crescendo with the creation of the Federal Department of Energy, and still sets the stage for a future energy catastrophe. Energy economist David Glasner concludes that the energy crisis is the product of political, bureaucratic, and business interests that benefit from government restrictions on energy markets.

"A healthy antidote for the too-widely shared view that price controls for oil and natural gas can protect the economy, the nation's security, or the consumer."
–GRANT P. THOMPSON
Conservation Foundation

"A superb job of describing and analyzing the economics and politics of energy regulation."
–*JOURNAL OF ECONOMIC LITERATURE*

297 pages, cloth 0-88410-953-4, $34.95
paper, 0-88410-954-2, $12.95

Available at better bookstores, or from
PACIFIC INSTITUTE, 177 Post Street,
Dept. A, San Francisco, CA 94108

PACIFIC INSTITUTE
FOR PUBLIC POLICY RESEARCH

ySTAT™

For IBM Personal Computers

Integrates SPREADSHEET + STATISTICS + FORECASTING + GRAPHICS and is Easy to Use.

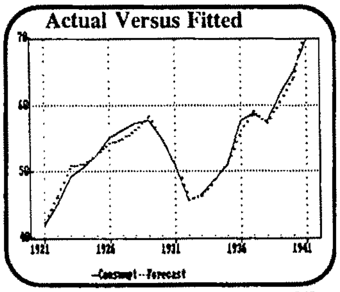
- o **ySTAT is the only full-fledged spreadsheet statistical program** - a fast, powerful, easy-to-use stand-alone program.
- o **Menu-driven format as in Lotus 1-2-3-** offers many functions that are not available even in 1-2-3, such as vector formulas, lagged variables, dummy variables, and row-column transposition.
- o **You can read, enter, edit, copy, move, and sort the data** as well as group and partition the data for analysis right on the spreadsheet. A total of 40 function keys for one-keystroke operations.
- o **Ten statistical function keys:** to tally sums and means; subtract variable means or seasonal means; standardize the data; aggregate seasonal data; generate seasonal indicators; perform moving averages, linear trend forecasts, and so on.
- o **Automatic recording feature** for programming macro keys, which with one keystroke, will replicate an entire sequence of analyses. You can save macros for future uses.
- o **ySTAT reads Lotus' files** and any free-format or fixed-format textfile generated by other PC programs or mainframe. The data can be numeric, alphabetic, including missing values.
- o **No disk swapping:** ySTAT is compactly written in C - the entire program resides on one disk.
- o **Descriptive statistics,** frequency distribution, correlation, crosstabs, analysis of variance.
- o **Multiple regression** (OLS), standardized regression, weighted least squares, two-stage least squares, polynomial regression, and Cochrane-Orcutt method. Including **regression diagnostics.** (see sample output at right)
- o Pooling of time-series and cross-sectional data.
- o **Two-sample Difference of means test, nonparametric tests** (Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, Mann-Whitney test, Wilcoxon matched-pairs signed-ranks test) and Spearmans's and Kendall's **rank-order correlations.**
- o **Time-series models:** autoregressive model, moving average model, and **Box-Jenkins model.**
- o **Nonseasonal forecasts:** linear trend, quadratic trend, polynomial model, sinusoidal model, and simple, double and triple exponential smoothings.
- o **Seasonal forecasts:** exponential smoothings with seasonal indicators or trigonometric functions; Winter's additive and multiplicative procedures.

SYSTEM REQUIREMENTS: IBM PC, XT, AT, or compatibles; a math co-processor is optional. Also available for Hewlett-Packard Portable Plus.

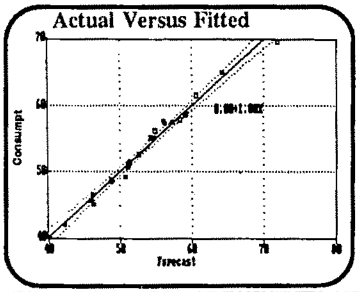
PRICE: $395. Including one free update.

```
                    OLS  -- DEPENDENT VARIABLE: Consumpt

  RIGHT-HAND          ESTIMATED       STANDARD         T_STATISTIC        PROB.
   VARIABLE          COEFFICIENT        ERROR

 1    Profits        0.192934381   (    0.09121)   T=    2.11527         0.049
 2        P-1        0.089884898   (    0.09065)   T=    0.99158         0.335
 3        W+W'       0.796218750   (    0.03994)   T=   19.93342         0.000
 4  Constant        16.236600272   (    1.30270)   T=   12.46382         0.000

  SAMPLE SIZE(  1 to   21) =              21           (DF=17)
  SUM OF SQUARED RESIDUALS =              17.879449
           VARIANCE (MSE) =               1.051732
 STANDARD ERROR (ROOT MSE) =              1.025540
                R-SQUARED =               0.981008
       ADJUSTED R-SQUARED =               0.977657
   F-STATISTIC(   3,   17) =            292.707595 (p=0.0004)
         SUM OF RESIDUALS =               0.000000
  DURBIN-WATSON STATISTIC =               1.367474

                             Source | SUM SQ    | DF |  MEAN SQ
                   |----------------|-----------|----|----------|
                   |Due to Regression|  941.430 |   3|   313.810|
                   |        Residual |   17.879 |  17|     1.052|
                   |           Total |  959.309 |  20|    47.965|
                   |----------------|-----------|----|----------|

RESIDUAL ANALYSIS - Mean: 0.000   Adj. RMSE: 0.946   Mean Abs. % Err: 1.283

                           AUTOCORRELATIONS

LAG   COEF    T-VAL   +-----+------+-----+-----+---0---+-----+------+-----+----+-  LAG
 1  0.181     0.83    |                            |>>>>>>>>>>                  |   1
 2 -0.059    -0.27    |                         <<<|                            |   2
 3 -0.033    -0.15    |                          <<|                            |   3
 4 -0.157    -0.69    |                    <<<<<<<<|                            |   4
 5 -0.411    -1.81    |<<<<<<<<<<<<<<<<<<<<<<<<<<<<|                            |   5
                      +-----+------+-----+-----+---0---+-----+------+-----+----+-  LAG
   Ljung-Box statistic (chi-square 4 DF): 6.698 (p=0.1527)

                        PARTIAL AUTOCORRELATION

LAG   COEF    T-VAL   +-----+------+-----+-----+---0---+-----+------+-----+----+-  LAG
 1  0.181     0.83    |                            |>>>>>>>>>>                  |   1
 2 -0.095    -0.44    |                      <<<<<<|                            |   2
 3 -0.004    -0.02    |                            |                            |   3
 4 -0.162    -0.74    |                  <<<<<<<<<<|                            |   4
 5 -0.378    -1.73    |<<<<<<<<<<<<<<<<<<<<<<<<<<<<|                            |   5
                      +-----+------+-----+-----+---0---+-----+------+-----+----+-  LAG

                  ACTUAL versus FITTED VALUES and RESIDUALS

SEQ    Actual    Fitted   min=41.900                   71.873=max   Error
        (*)       (+)     +-----+-----+-----+-----+-----+-----+-+              +---+---+
 1     41.90     42.22    |@                                    |    -0.324   |  <:    |
 2     45.00     46.25    |   *+                                |    -1.250   |<<<:    |
 3     49.20     50.77    |       *  +                          |    -1.566   |<<<:    |
 4     50.60     51.09    |          @                          |    -0.494   |  <:    |
 5     52.60     52.59    |            @                        |     0.008   |   :>   |
 6     55.10     54.23    |              +*                     |     0.869   |   :>>  |
 7     56.20     54.86    |              +  *                   |     1.338   |   :>>> |
 8     57.30     56.25    |                +*                   |     1.055   |   :>>  |
 9     57.80     58.39    |                 *+                  |    -0.589   | <<:    |
10     55.00     54.72    |              @                      |     0.282   |   :>   |
11     50.90     51.13    |          @                          |    -0.230   |  <:    |
12     45.60     45.92    |   *+                                |    -0.322   |  <:    |
13     46.50     46.18    |    @                                |     0.322   |   :>   |
14     48.70     48.76    |       @                             |    -0.058   |  <:    |
15     51.30     51.33    |          @                          |    -0.035   |  <:    |
16     57.70     56.08    |               +*                    |     1.616   |   :>>> |
17     58.70     59.14    |                 *+                  |    -0.436   |  <:    |
18     57.50     57.29    |                @                    |     0.210   |   :>   |
19     61.60     60.61    |                   +  *              |     0.989   |   :>>  |
20     65.00     64.21    |                       +*            |     0.785   |   :>>  |
21     69.70     71.87    |                             *     +|    -2.173   |<<<<:   |
        (*)       (+)     +-----+-----+-----+-----+-----+-----+-+              +---+---+

        LIST OF OUTLIERS - STUDENTIZED RESIDUALS GREATER THAN 1.5

 SEQ     Actual   Forecast   Residual   Std. Err  Studentized Probability
                                                   Residuals     DF=17
 ----    ------   --------   --------   --------  ----------  ----------
    3    49.20     50.77     -1.566      0.967     -1.619       0.124
   16    57.70     56.08      1.616      0.958      1.687       0.110
   21    69.70     71.87     -2.173      0.780     -2.787       0.013
 ----    ------   --------   --------   --------  ----------  ----------

                     Covariance-Variance Matrix

                   Profits        P-1         W+W'     Constant

  1   Profits      0.0083      -0.0053      -0.0013     0.0005
  2       P-1     -0.0053       0.0082      -0.0007    -0.0177
  3       W+W'    -0.0013      -0.0007       0.0016    -0.0329
  4  Constant      0.0005      -0.0177      -0.0329     1.6970
[Note: this sample output shows some of the options for regression diagnosis.]
```

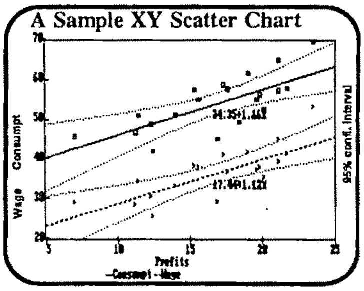
Actual Versus Fitted



Actual Versus Fitted



A Sample XY Scatter Chart

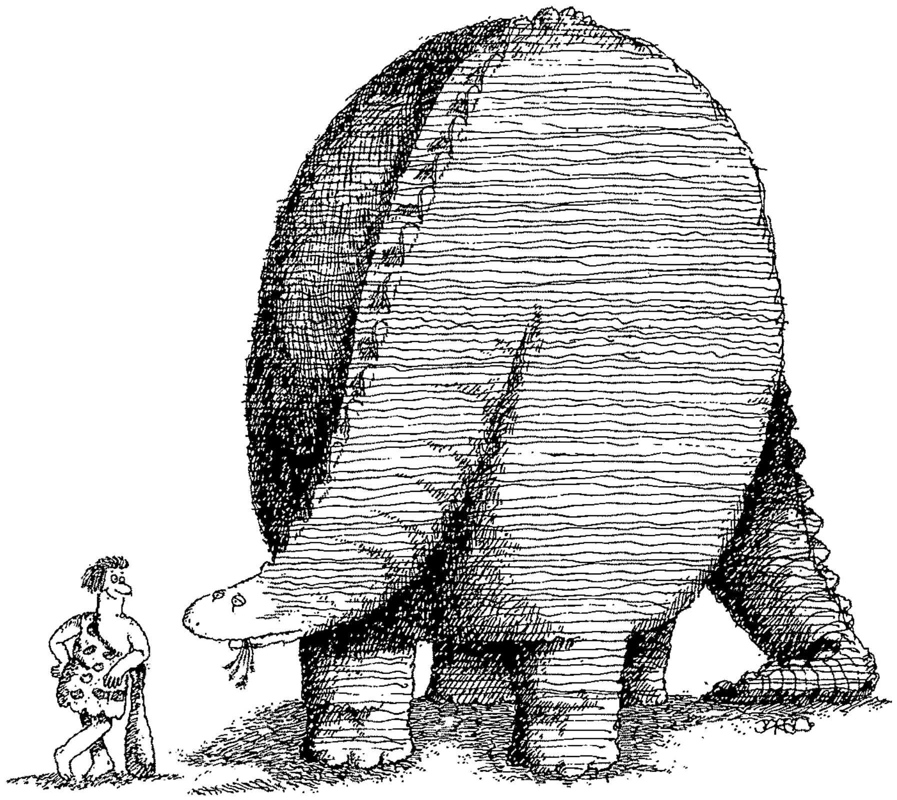


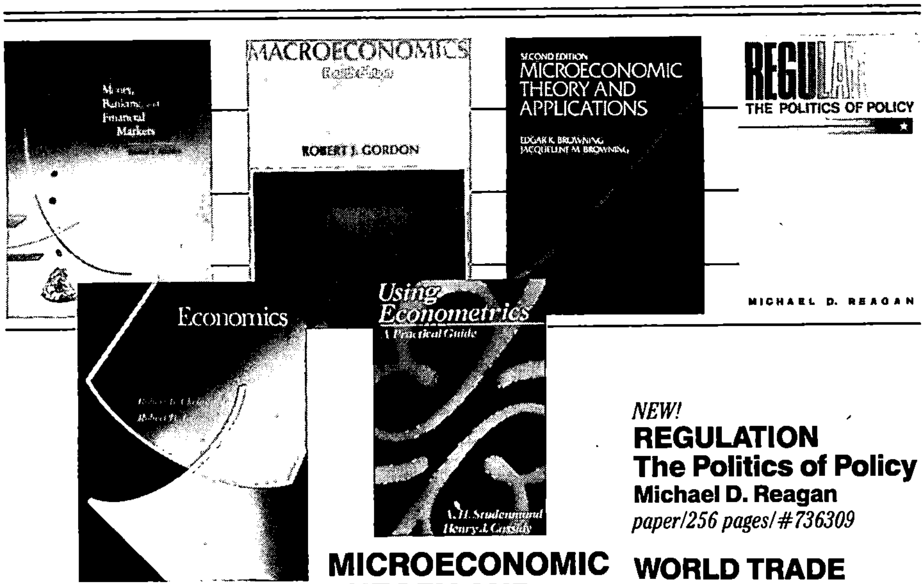
MING TELECOMPUTING INC.

23 Oak Meadow Road, P.O. Box 101, Lincoln Center, MA 01773, U.S.A. (617) 259-0391

Order: ( ) ySTAT Package for $395.
( ) Trial Disk and information/sample output for $5.
( ) Information/sample output.
System: ( ) IBM PC ( ) XT ( ) AT. Memory ______K
( ) Other system ____________________
Options: ( ) with 8087 or 80287 math co-processor.
( ) IBM CGA ( ) EGA ( ) Hercules Mono. Card
( ) Other card ____________________

Payment: ( ) Check or money order.
( ) U.S. university or governmental purchase order.
( ) Visa. ( ) MC. Card No. ____________________
Name ____________________ Expir. date ___/___
Address ____________________
____________________ Telephone: ( ) __________





NEW!

USING ECONOMETRICS
A Practical Guide
A.H. Studenmund
Occidental College
Henry J. Cassidy
Federal Home Loan Mortgage Corp.

This innovative new introductory text provides students with practical advice on how to use the tools of econometrics in real-world situations. Single-equation linear regression analysis is introduced to students through extensive and relevant examples. The text stresses an intuitive approach and holds the use of formal mathematical derivations to a minimum. In addition, specification issues—including uses and abuses of econometrics—are given more attention here than is usual in comparable texts.
cloth/528 pages/#820105/ Instructor's Manual

MICROECONOMIC THEORY AND APPLICATIONS
Second Edition
Edgar K. Browning
Jacquelene M. Browning
both of Texas A&M University

Hundreds of instructors and thousands of students benefited from the Brownings' formula for teaching microeconomics in their First Edition. Still focusing on four core areas—the theories of consumer choice, competitive markets, monopolies, and input markets—the Second Edition features new applications, such as automobile import quotas and peak-load pricing, and a revised chapter on oligopoly, with new material on game theory. Tables, graphs, chapter summaries, and over 100 new end-of-chapter questions lend ample learning support.
cloth/637 pages/#112356/ 1986/Instructor's Manual, Study Guide and Workbook

NEW!

REGULATION
The Politics of Policy
Michael D. Reagan
paper/256 pages/#736309

WORLD TRADE AND PAYMENTS
An Introduction
Fourth Edition
Richard E. Caves and Ronald W. Jones
cloth/537 pages/1985/#132276

INTERNATIONAL TRADE AND FINANCE
Readings
Third Edition
Robert E. Baldwin and J. David Richardson
paper/550 pages/#079278/1986

THE ECONOMICS OF ANTITRUST
Cases and Analysis
Don E. Waldman
paper/320 pages/#917915/1986

REGULATORY REFORM
What Actually Happened
Leonard W. Weiss and Michael W. Klass
paper/316 pages/#928984/1986

LITTLE, BROWN AND COMPANY

College Division • 34 Beacon Street • Boston, MA 02108-1493

BALLINGER

MONETARISM AND LIBERALIZATION
The Chilean Experiment

Sebastian Edwards and
Alejandra C. Edwards

An in-depth examination of the Pinochet government's experiment in free-market policies from promising beginnings in 1973 to ultimate failure in 1982, *Monetarism and Liberalization* looks at what went wrong in Chile. Why did liberalization and stabilization programs initially succeed in dramatically reducing inflation and raising real output only to collapse nine years later? This analysis illuminates some of the more puzzling aspects of the Chilean experiment and extracts vital lessons.

January 1987 0-88730-105-3 256 pages $29.95, cloth

THE ECONOMICS OF TELECOMMUNICATIONS
Theory and Policy

John T. Wenders

With the breakup of AT&T and the onset of "telephone wars" between AT&T and companies like MCI and Sprint, telephone services as we knew it changed for good. In *The Economics of Telecommunications,* John Wenders, a consultant to AT&T before its divestiture, explains the economic and political forces behind the breakup and looks at the future of the telecommunications industry. He develops a framework for understanding the economic basis for regulation of the industry, argues that deregulation will be beneficial, and then calls for rapid changes in the industry.

January 1987 0-88730-119-3
296 pages $34.95, cloth

THE PARADOX OF POVERTY
A Reappraisal of Economic Development Policy

Paul Steidlmeier

Questioning the very foundations of the world economic order, Steidlmeier provides a much-needed reexamination of development policy. Drawing on economic and technical analysis as well as sociocultural factors, Steidlmeier reveals the proper role of economic development and its power to effect social change. This unique and integrated approach focuses on five variables: population, environment, technology, power and social organization, and culture.

January 1987 0-88730-184-3
344 pages $29.95, cloth

MANAGED TRADE
The New Competition Between Nations

Raymond J. Waldmann

The United States, which espouses a free trade policy worldwide, breaks its own free trade rules. And it is not alone. Surprised?

Revealing what is *really* going on in world trade today, *Managed Trade* proposes solutions to the discrepancies between policy and practice. Waldmann advocates revitalizing international trade to combine consistent business practices with economic efficiency.

November 1986 0-88730-143-6
224 pages $29.95, cloth

INNOVATION AND MARKET STRUCTURE
Lessons from the Computer and Semiconductor Industries

Nancy S. Dorfman

In *Innovation and Market Structure* Nancy S. Dorfman distinguishes the roles of small, new enterprises from those of large, established firms in driving the wheels of innovation and technological progress. Showing that the rules for successful innovation are *not* the same for all kinds of businesses, Dorfman examines the opportunities and incentives for small, new firms to join the innovation process and share in the fruits of technological progress.

January 1987 0-88730-185-1
280 pages $32.00, cloth

BALLINGER

YES! Please send me:
___**Monetarism and Liberalization** (6610729) cloth, $29.95
___**Managed Trade** (6611065) cloth, $29.95
___**Innovation and Market Structure** (6611560) cloth, $32.00
___**The Economics of Telecommunications** (6610844) cloth, $34.95
___**The Paradox of Poverty** (6611057) cloth, $29.95

□ Payment enclosed □ Bill me
Charge my □ MC □ VISA □ AMX

Card no.____________________Exp. date______
Signature__________________________
Send to:__________________________
____
__________________________Zip______

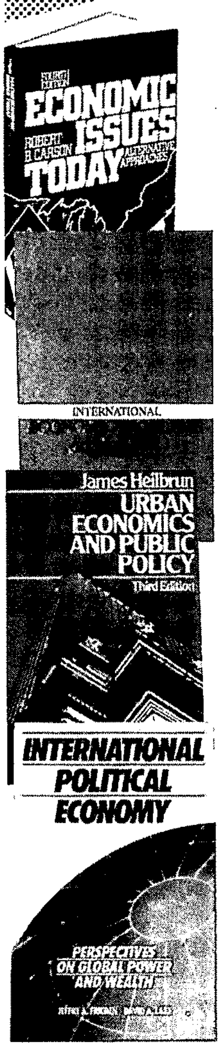
Harper & Row
Order Department
2350 Virginia Avenue
Hagerstown, MD 21740

(800) 638-3030

My state sales tax $______
Postage/handling ($1.50/bk)* $______
***Prepaid orders are postage free!**
TOTAL $______

Prices subject to change. All orders subject to credit approval. U.S. funds only. If you order by phone, tell the operator your order code is **AAER387**

St. Martin's Press 1987 economics titles ...addressing the major issues



Economic Issues Today
ALTERNATIVE APPROACHES
Fourth Edition

ROBERT B. CARSON, State University of New York, College at Oneonta

An objective examination of sixteen current economic issues from three different ideological perspectives—conservative, liberal, and radical.
Paperbound 386 pages November 1986

Also available in two separate volumes—
Macroeconomic Issues Today and
Microeconomic Issues Today

International Economic Problems and Policies

H. PETER GRAY, Rutgers University

An issue-oriented introduction to international economic phenomena and their influence on domestic policies that examines both the macro issues of commercial policy (trade) and the mechanics of international payments (finance).
486 pages February 1987

Urban Economics and Public Policy
Third Edition

JAMES HEILBRUN, Fordham University

A thorough introduction to the study of urban economics that offers a balanced treatment of urban economic theory and public policy issues.
506 pages February 1987

International Political Economy
PERSPECTIVES ON GLOBAL POWER AND WEALTH

JEFFREY A. FRIEDEN and DAVID A. LAKE,
both of University of California, Los Angeles

A comprehensive reader that examines the three major theoretical perspectives of international political economy—Liberalism, Marxism, and Realism—focusing on international investment, money, and trade.
Paperbound 420 pages November 1986

To request an examination copy of any of these titles, please write us on your college letterhead, specifying your course title, current text, and approximate enrollment.
Send your request to:

ST. MARTIN'S PRESS
Department JR, 175 Fifth Avenue, New York, NY 10010 SMP

New and Noteworthy

New **Macroeconomics and Finance**

Essays in Honor of Franco Modigliani

edited by Rudiger Dornbusch, Stanley Fischer, and John Bossons

"Modigliani and his coworkers give, in brief and objective form, what we all need: a survey of where macroeconomics stands today—a rating of the different schools and paradigms."
—Paul A. Samuelson, MIT $40.00

New **Essays in Economics**

Volume 1: Macroeconomics

James Tobin

These 28 essays cover Tobin's work from the early 1940s to 1970 in macroeconomic theory, economic growth, and money and finance.
$60.00

Superfairness

Applications and Theory

William J. Baumol

"William Baumol's work gives economists a theory of equity that can be applied with the rigor, and taught with the clarity, of Paretian efficiency theory. Everyone who teaches or uses welfare economics will benefit from reading this book."—Duncan K. Foley, Barnard College of Columbia University $20.00

New **Energy**

Markets and Regulation • Essays in Honor of M.A. Adelman

edited by Richard L. Gordon, Henry D. Jacoby, and Martin B. Zimmerman

This valuable survey of current thinking on energy economics covers nearly every aspect of the energy sector, including both international and domestic markets in oil and coal and the particular U.S. conditions in natural gas and nuclear power. The authors deal with resource estimation and energy supply and demand, and environmental control. $37.50

Regulation of Economic Activity Series
Richard Schmalensee, editor

New **Public Regulation**

New Perspectives on Institutions and Policies

edited by Elizabeth E. Bailey

These eleven original essays propose and demonstrate a wide variety of new techniques for assessing regulatory performance—general equilibrium theory, laboratory methods, the integration of rational expectations and game theory methods with regulations, cost-functions techniques, and probabilistic risk analysis—to reveal new empirical data from a range of regulated industries such as agriculture, natural gas, nuclear power, airlines, health and environmental. $42.50

New **The Economics of Public Utility Regulation**

Michael A. Crew and Paul R. Kleindorfer

The Economics of Public Utility Regulation surveys the large literature on the regulation of public utilities and provides industry studies with specific applications of the more general theories. $22.50

Timely Economic Analyses

from

Lexington Books

The Deficit Problem in Perspective
Richard J. Cebula, Emory University
What are the economic effects of the federal budget deficit? How useful are policies like the Balanced Budget Amendment and the Gramm-Rudman-Hollings Act? This up-to-the-minute textbook comprehensively explains the questions surrounding the deficit, and is essential reading for students of economics, public finance, and money and banking.
ISBN: 0-669-14303-0 128p. Feb. $12.00 (tent., pbk)

The Fairness of Markets: A Search for Justice in a Free Society
Richard B. McKenzie, Clemson University
Are society's ills best solved by government economic policies or by free market forces? This provocative book examines a wide range of federal policies and draws some startling conclusions about their effects on social justice. The thoughtful analysis provides valuable guidance on the economic direction the United States must take in the post-Reagan era.
ISBN: 0-669-14801-6 288p. March $24.95

The General Trading Company: Concept and Strategy
Dong-Sung Cho, Seoul National University
An in-depth look at the unique advantages of a new form of international business: the general trading company. The author compares four types of international business organizations to identify the best strategies for international trade.
ISBN: 0-669-14296-4/176p./$25.00

China's Development Strategies and Foreign Trade
James T.H. Tsao, Georgetown University
An impressive analysis of China's economic development strategies in the past, present, and future. The author looks closely at foreign trade and links current economic reforms to changes in the U.S.-Chinese trade relationship.
ISBN: 0-669-13954-8/224p./Feb./$32.00 (tent.)

The Export-Import Bank at Fifty: The International Environment and the Institution's Role
Rita M. Rodriguez, Export-Import Bank
A careful assessment of the achievements of the Export-Import Bank and a look at the trends that will shape its future role. The contributors include a number of prominent economists, government officials, experts on international development, and academics.
ISBN: 0-669-14828-8/224p./$29.00

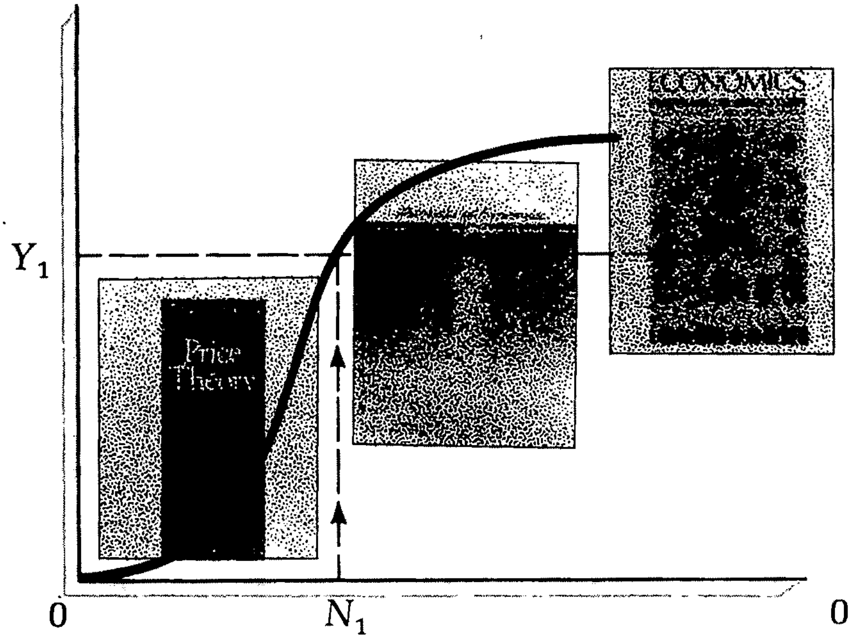
DC Heath
A **Raytheon** Company

Federal Budget Deficits: An Economic Analysis
Richard J. Cebula, Emory University
Is the federal budget deficit a serious economic crisis? Is there a cure? This book offers a rigorous explication of the effects of the deficit and the measures used to correct it.
ISBN: 0-669-11095/192p./$29.00

The Sunkist Case: A Study in Legal-Economic Analysis
Willard F. Mueller, Peter G. Helmberger, University of Wisconsin, and *Thomas W. Patterson,* Susman, Godfrey and McGowan
A thought-provoking analysis of the legal and economic ramifications of the Sunkist monopoly case. The authors refute the FTC charges, show how antitrust laws can endanger agricultural monopolies, and highlight how pricing strategies function in the marketplace.
ISBN: 0-669-15189-0/320p./Apr./$34.95 (tent.)

Social Security: A Critique of Radical Reform Proposals
Charles W. Meyer, Iowa State University
Should social security be modified or phased out? This book presents stimulating dissections of the Ferrara phase-out plan by a number of distinguished contributors.
ISBN: 0-669-14518/176p./Jan./$32.00 (tent.)

Lexington Books/D.C. Heath
125 Spring Street
Lexington, Massachusetts 02173
(800) 235-3565

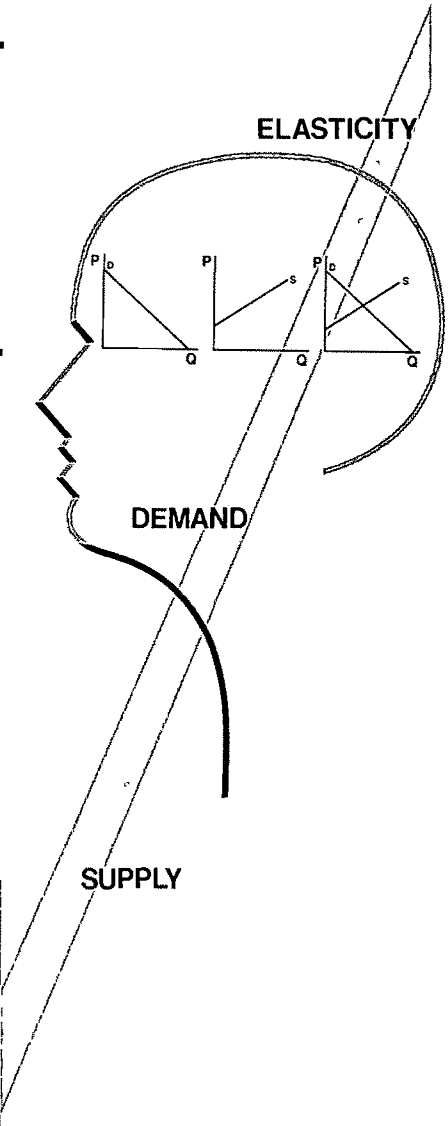

Y1
0
N1
0
Price Theory
ECONOMICS

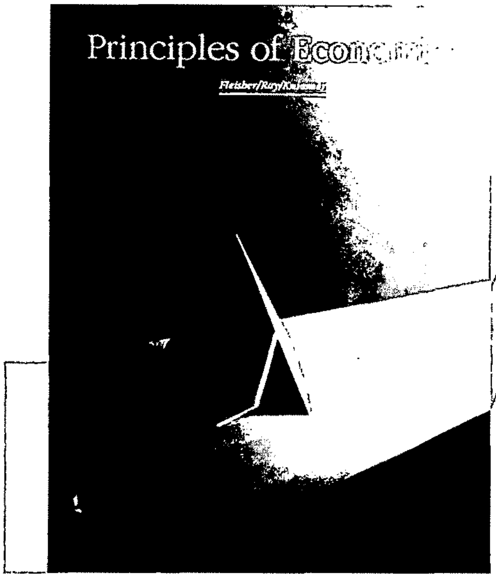
"The writing style will make economics accessible to students of every background."

▲Gerard Russo
University of Wisconsin,
Madison

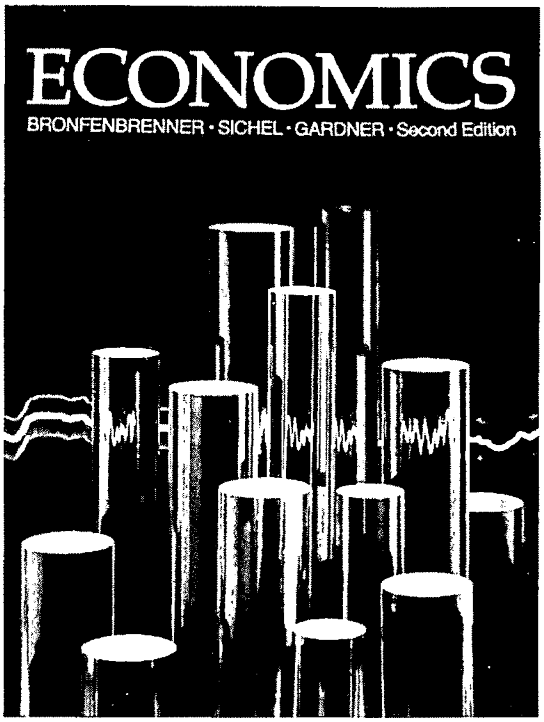
"The text is rigorous but accessible . . . due to the authors' mental precision and writing skill."

▲Jerome L. McElroy
Saint Mary's College





wcb WM. C. BROWN
PUBLISHERS
2460 Kerper Boulevard P.O. Box 539
Dubuque, IA 52004-0539



THE RATIONAL CHOICE IS THE NATIONAL CHOICE

The first edition of Bronfenbrenner/Sichel/Gardner—the rational choice—brought the economics of the '80s into the mainstream of the principles course. Now, in its second edition, *Economics* provides an even stronger presentation of aggregate demand and aggregate supply, increases emphasis on applied microeconomics, and discusses current economic developments.

ECONOMICS, Second Edition

Martin Bronfenbrenner, Aoyama Gakuin University, Japan
Werner Sichel and **Wayland Gardner,** both of Western Michigan University

Complete hardcover edition: About 847 pages

Two-Volume paperback edition:
MACROECONOMICS About 451 pages • **MICROECONOMICS** About 572 pages

Complete support package available • Just published

For adoption consideration, request an examination package from your regional Houghton Mifflin office.

Houghton Mifflin Company 13400 Midway Road, Dallas, TX 75244-5165/
1900 South Batavia Avenue, Geneva, IL 60134/989 Lenox Drive, Lawrenceville, NJ 08648/
925 East Meadow Drive, Palo Alto, CA 94303

FAST • EASY TO USE • POWERFUL

THE GAUSS™

MATHEMATICAL AND STATISTICAL SYSTEM

Written by Lee E. Edlefsen and Samuel D. Jones

FOR IBM PC/XT/AT AND COMPATIBLES

GAUSS is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines **speed, power, and ease of use** in one amazing program.

GAUSS allows you to do essentially **anything you can do with a mainframe** statistical package — and a lot more.

Personal computers are **friendly, convenient, and inexpensive.** So is **GAUSS. GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS,** you may never use a mainframe again.

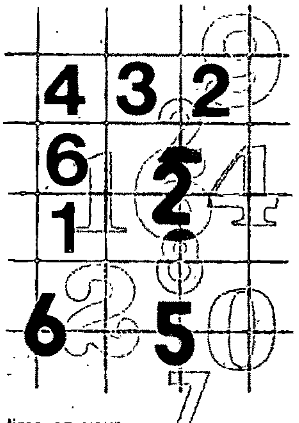
GAUSS comes with programs written in its matrix programming language that allow you to do most **statistical procedures,** including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is **no limit** on the number of observations.

GAUSS will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the **maximum likelihood** estimates of a binary logit model, with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

GAUSS allows you to do **complicated statistical procedures** that you would never imagine trying on a mainframe. It is **easy to program** almost any routine, and **GAUSS** is so **fast** that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

GAUSS is an **excellent teaching tool.** It makes programming easy and allows students to focus on concepts and techniques.



If you can write it mathematically, you can write it in **GAUSS.** Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

GAUSS is **10-15 times faster** than other programs that use the 8087, and **15-100 times faster** than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional **powerful numerical operators and functions** — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

GAUSS has **state-of-the-art numerical routines** and **random number generators.**

GAUSS is **extremely accurate.** It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coeficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

GAUSS, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing simulations.

GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.

GAUSS Programming Language **$200.00**
(includes Applications Module 1) plus shipping/handling

GAUSS Mathematical and Statistical System **$300.00**
(includes Language, Applications Modules 1-11) plus shipping/handling

GAUSS is offered with a **30 day money-back guarantee.** GAUSS is not copy protected. Requires an IBM PC/XT/AT or compatible with 8087/80287, 2 DS/DD disk drives and/or hard disk, 320-640K RAM, DOS 2.1 or above. Graphics require IBM Color Graphics Adaptor or equivalent, with suitable monitor.

Call or Write

APTECH SYSTEMS, INC.
P.O. Box 6487, Kent, WA 98064
(206) 631-6679

10. 11. 85

The American Economic Review

PAPERS AND PROCEEDINGS

OF THE

Ninety-Ninth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New Orleans, Louisiana, December 28–30, 1986

Program Arranged by Gary S. Becker

Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John

MAY 1987

THE AMERICAN ECONOMIC ASSOCIATION

• Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

• *THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1987, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

$38.50	if annual income is $30,000 or less;
$46.20	if annual income is above $30,000, but no more than $40,000;
$53.90	if annual income is above $40,000.

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), $110 a year; individuals, $72.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $12.00 to the annual rates above to cover extra postage.

• Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

• Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

THE AMERICAN ECONOMIC REVIEW

VOL. 77 NO. 2 MAY 1987

PAPERS AND PROCEEDINGS

. OF THE

Ninety-Ninth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New Orleans, Louisiana

December 28–30, 1986

Program Arranged by Gary S. Becker

Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John

CONTENTS

PAPERS

Richard T. Ely Lecture

Roundtable on Teaching Undergraduate Courses in Quantitative Methods

Finance and Economics

Dynamic Models of Taxation

New Theories of Economic Growth

The Family

Multiple Equilibria in Macroeconomics

Deterrence and Enforcement of Laws

Time-Series Analyses of Macroeconomic Flucuations

The Contribution of Keynes After 50 Years

Economic Status of Minorities

The International Debt Crisis

Game Theory and Industrial Organization

Reforming the International Monetary System

Continuing Black Poverty

The Economic Analysis of Taxpayer Compliance

Gender Differences in Behavior at Home and at Work

Lessons from Health Economics

Stopping High Inflation

Economics and Politics

The Economics of Discrimination Thirty Years Later

Real Business Cycle Theory: What Does It Explain?

Arbitration and the Negotiation Process

PROCEEDINGS

Reports

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-ninth annual meeting of the American Economic Association. The *Proceedings* record the business activities of the Association in 1986; the annual membership meeting; and the March and December meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise fifty-eight contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. We would like to take this opportunity to answer a number of commonly asked questions about the *Papers*.

Who chooses the authors? About a year in advance, the Association's President-elect, acting as a program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* is published annually in the Notes section of the December issue of the *AER*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing twenty-three sessions, although a total of ninety-two sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

Are discussants' comments published? There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect decided to publish no comments, given the difficulty and the invidious task of choosing. He has arranged instead that the names and affiliations of commentators be printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

What standards must the papers meet? The guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, very rigid deadlines must be met and there is not time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long, or does not satisfy other specifications.

This year, most authors cooperated very nicely. We thank them for making our lives easier. To those who failed to follow the guidelines, we suggest a reading of *Proverbs* 13:18.

HARVEY S. ROSEN
WILMA ST. JOHN

RICHARD T. ELY LECTURE

The Law and Economics Movement

By RICHARD A. POSNER*

In the last thirty years, the scope of economics has expanded dramatically beyond its traditional domain of explicit market transactions.[1] Today there is an economic theory of property rights, of corporate and other organizations, of government and politics, of education, of the family, of crime and punishment, of anthropology, of history, of information, of racial and sexual discrimination, of privacy, even of the behavior of animals—and, overlapping all these but the last, of law.[2]

Some economists oppose this expansion, in whole or (more commonly) in part.[3] There are a number of bad reasons, all I think closely related, for such opposition, and one slightly better one.

1) One bad reason is the idea that economics *means* the study of markets, so that nonmarket behavior is simply outside its scope. This type of argument owes nothing really to economics, but instead reflects a common misconception about language—more specifically a failure to distinguish among three different types of word or concept. The first type, illustrated by the term "marginal cost," is purely conceptual. The term is rigorously and unambiguously defined by reference to other concepts, just as numbers are; but (again like numbers) there is no observable object in the real world that it names. (Try finding a firm's marginal costs on its books of account!) The second type of word, illustrated by "rabbit," refers to a set of real-world objects. Few such words are purely referential; one can speak of a pink rabbit or a rabbit the size of a man without misusing the word, even though one is no longer using it to describe anything that exists. Nevertheless, the referential function dominates. Finally, there are words like "law," "religion," "literature"—and "economics"—which are neither conceptual nor referential. Such words resist all efforts at definition. They have, in fact, no fixed meaning, and their dictionary definitions are circular. They can be used but not defined.[4]

One cannot say that economics is what economists do, because many noneconomists do economics. One cannot call economics the science of rational choice, either. The word "rational" lacks a clear definition; and, passing that difficulty, there can be noneconomic theories of rational choice, in which few predictions of ordinary economics may hold; for example, because the theory assumes that people's preferences are unstable.

There can also be nonrational economic theories; an example is the type of survival theory in industrial organization in which firms that randomly hit on methods of lowering their costs expand vis-à-vis their rivals; another example is Marxism. One cannot call economics the study of markets either, not only because that characterization resolves the question of the domain of eco-

*Judge, U.S. Court of Appeals for the Seventh Circuit; Senior Lecturer, University of Chicago Law School, 1111 E. 60th St., Chicago, IL 60637. I thank Gary Becker, Frank Easterbrook, William Landes, Geoffrey Miller, Richard Porter, George Stigler, Geoffrey Stone, and Alan Sykes for comments on a previous draft and Nir Yarden for research assistance.

[1] See, for example, Gary Becker (1976); Jack Hirshleifer (1985, p. 53); George Stigler (1984); Gerard Radnitzky and Peter Bernholz (1986).

[2] For a recent conspectus of economic analysis of law, see my book (1986).

[3] See, for example, Ronald Coase (1978). Coase is of course a leading figure in the economics of property rights, so his opposition is far from total.

[4] For an excellent discussion, see John Ellis (1974, ch. 2).

nomics by an arbitrary definitional stop but also because other disciplines, notably sociology, anthropology, and psychology, also study markets. About the best one can say is that there is an open-ended set of concepts (such concepts as perfect competition, utility maximization, equilibrium, marginal cost, consumers' surplus, elasticity of demand, and opportunity cost), most of which are derived from a common set of assumptions about individual behavior and can be used to make predictions about social behavior; and that when used in sufficient density these concepts make a work of scholarship "economic" regardless of its subject matter or its author's degree. When economics is "defined" in this way, there is nothing that makes the study of marriage and divorce less suitable a priori for economics than the study of the automobile industry or the inflation rate.

2) The "extension" of economics from market to nonmarket behavior is sometimes thought to be premature until the main problems in the study of explicit markets have been solved. How can economists hope to explain the divorce rate when they can't explain behavior under oligopoly? But this rhetorical question is just a variation on the first point, that economics has a fixed subject matter, a predefined domain. The tools of economics may be no good for solving a number of important problems in understanding explicit markets; that is no reason to keep hitting one's head against the wall. Economics does not have a predestined mission to dispel all the mysteries of the market. Maybe it will do better with some types of nonmarket behavior than with some types of market behavior.

3) Next is the idea that to do economics in fields that have their own scholarly traditions, such as history or law, an economist must master so much noneconomic learning that his total educational investment will be disproportionate to the likely fruits of "interdisciplinary" research; hence economists should steer clear of these fields. Besides disregarding the possibility of collaboration between economists and practitioners of other disciplines, this argument assumes that economics means something done by people with a Ph.D. in economics. It may be easier for an anthropologist to learn economics than for an economist to learn anthropology. Maybe the fraction of one's training in economics that is irrelevant to the economic analysis of anthropological phenomena is larger than the fraction of anthropological training that is irrelevant; or maybe economic theory is more compact than the body of knowledge we call anthropology. (It probably is easier to learn economics well than to learn Chinese well.) Or it might simply be (this has happened in law and economics) that a given anthropologist had more of a knack for economics than a given economist had a knack for anthropology. It is only by defining economics, in rather a medieval way, as the work done by members of a particular guild (the guild of economics Ph.D.s) that one will be led to conclude that if the economics of law is done by lawyers, or the economics of history by historians, it cannot be "real" economics. The emergence of nonmarket economics may have resulted in a vast but unrecognized increase in the number of economists!

The idea that nonmarket economics is somehow peripheral to economics is connected with the fact that there has been little fruitful analysis of explicit markets besides economics, though admirers of Max Weber's analysis of the role of Protestantism in the rise of capitalism may want to challenge this assertion. Almost by default, explicit markets became thought of as the natural subject matter of economics. But the fact that other areas of social behavior, such as law, have been extensively studied from other angles than the economic is no reason for concluding that these areas cannot be studied profitably with the tools of modern economic theory.

4) Still another bad reason for hostility to nonmarket economics is fear that it will bring economics into disrepute by associating the economist with politically and morally distasteful, bizarre, or controversial practices (such as capital punishment, polygamy, or slavery before the Civil War) and proposals—whether specific policy proposals such as education vouchers, or the idea, which is basic to nonmarket economics,

that human beings are rational maximizers throughout the whole, or at least a very broad, range of their social interactions. If economics becomes associated with highly sensitive topics, it may lose some of the appearance of scientific objectivity that economists have worked so hard to cultivate in the face of obvious difficulties including the fact that much of traditional microeconomics and macroeconomics is already politically and ethically controversial, as is evident from current debates over free trade, deregulation, and deficit spending. But this complaint, too, is part of the fallacious idea that there is a fixed domain for economics. If there were, it would be natural to recoil from economic ventures at once peripheral and controversial. But if I am right that there is no fixed, preordained, or natural domain for economics—that politics, punishment, and exploitation are, at least a priori, as appropriate subjects for economics as the operation of the wheat market—then it is pusillanimous to counsel avoidance of particular topics because they happen to be politically or ethically (are these different?) controversial at the present time.

5) A slightly better reason for questioning the expansion of economics beyond its traditional boundaries is skepticism that economic tools will work well in the new fields or that adequate data will be available in them to test economic hypotheses. Maybe these are domains where emotion dominates reason, and maybe economists can't say much about emotion. And explicit markets generate substantial quantitative data (prices, costs, output, employment, etc.), which greatly facilitate empirical research—though only a small fraction of economists actually do empirical research. These points suggest a functional as distinct from a definitional answer to the question of the appropriate bounds of economics: economics is the set of fruitful applications of economic theory. But a detailed survey of nonmarket economics is not necessary in order to make the point that the economic approach has been shown to be fruitful in dealing with such diverse nonmarket subjects as education, economic history, the causes of regulatory legislation, the behavior of nonprofit institutions, divorce, racial and sexual wage differentials, the incidence and control of crime, and (I shall argue) the common law rules governing property, torts, and contracts[5]—successful enough at any rate to establish nonmarket economics as a legitimate branch of economics, and to counsel at least a temporary suspension of disbelief by the skeptics and doubters. Indeed, so familiar have some of these areas of nonmarket economics become in recent years (for example, education viewed through the lens of human capital theory) that many young economists no longer think of them as being outside the traditional boundaries of economics. The distinction between "market" and "nonmarket" economics is fraying.

I

A

The particular area of nonmarket economics that I want to focus on is the economics of law, or "law and economics" as it is often called. Because of the enormous range of behavior regulated by the legal system, law and economics could be defined so broadly as to be virtually coextensive with economics. This would not be a useful definition. Yet to exclude bodies of law that regulate explicit markets—such as contract and property law, labor, antitrust and corporate law, public utility and common carrier regulation, and taxation—would be cripplingly narrow. But if these bodies *are* included, in what sense is law and economics a branch of *non*market economics? (I do not suggest that this is an important question; it may, indeed, be an argument for discarding an increasingly uninteresting distinction.)

As with any nonreferential, nonconceptual term, the only possible criterion for a definition of law and economics is utility—not accuracy. The purpose of carving out a separate field and calling it law and economics

[5] For a few examples see Becker (1981; 1975); Orley Ashenfelter and Albert Rees (1973); Robert Fogel and Stanley Engerman (1971); Isaac Ehrlich (1974); David Pyle (1983); Stigler (1971).

(or better, because clearer, "economics of law") is to identify the area of economic inquiry to which a substantial knowledge of law in both its doctrinal and institutional aspects is relevant. Many economic problems in such areas of law as taxation and labor do not require much legal knowledge to solve. Although taxes can be imposed only by laws, often the details of the tax law either are not relevant to the analyst, as where he is asking what the effect on charitable giving of reducing the marginal income tax rate is likely to be, or are transparent and unproblematic.

Similarly, in the field of labor, you can study the effects of unemployment insurance on unemployment without knowing a great deal about the state and federal laws governing unemployment insurance, though you must know something. But suppose you wanted to study the consequences of allowing the defendant in an employment discrimination case to deduct from the lost wages awarded the plaintiff (if the plaintiff succeeds in proving that he was fired because of race or sex or some other forbidden criterion), any unemployment benefits that the plaintiff might have received after being fired. You could not get far in such a study without knowing a fair amount of nonobvious employment discrimination law: Is there a uniform judicial rule on deduction or nondeduction of such benefits? Could the benefits be deducted but then be ordered paid to the state or the federal government rather than kept by the employer? Does the law insist that the employee who wants damages for employment discrimination search for work? How are those damages computed? The economics of law is the set of economic studies that build on a detailed knowledge of some area of law; whether the study is done by a "lawyer," an "economist," someone with both degrees, or a lawyer-economist team has little significance.

The law and economics movement has made progress in a number of areas of legal regulation of explicit markets. These include antitrust law, and the regulation of public utilities and common carriers; fraud and unfair competition; corporate bankruptcy, secured transactions, and other areas of commercial law; corporate law and securities regulation; and taxation, including state taxation of interstate commerce, an area that the courts regulate under the commerce clause of the Constitution.[6] In none of these areas is participation by economists, or (if we insist on guild distinctions) by economics-minded lawyers, particularly controversial any more, though some die-hard lawyers continue to resist the encroachments of economics and of course there is disagreement among economists over many particular issues; this is notable in antitrust. An area of legal regulation of explicit markets that is just beginning to ripen for economics is intellectual property, with special reference to copyrights and trademarks. Patents have long been an object of economic study.

The areas of law and economics about which economists and lawyers display considerable unease are the (sometimes arbitrarily classified as) nonmarket areas—crime, torts, and contracts; the environment; the family; the legislative and administrative processes; constitutional law; jurisprudence and legal process; legal history; primitive law; and so on. All the reasons that I gave at the outset for why some economists resist the extension of economics beyond its traditional domain of explicit market behavior coalesce in regard to these areas. And because they are also close to the heart of what lawyers think distinctive about law—of what they think makes it something more than a method of economic regulation—this branch of economic analysis of law dismays many lawyers. Furthermore, lawyers tend to have more rigid, stereotyped ideas of the boundaries of economics than economists do, in part because most lawyers are not aware of the extension (which is recent, though its roots go back to Adam Smith and Jeremy Bentham) of economics to nonmarket behavior. Indeed, a demarcation

[6] The work in these areas is summarized in my book (pts. 3–5 and ch. 26). It is of some interest to note that the economic analysis of secured financing is now dominated by economically inclined lawyers. See Robert Scott (1986) and references cited there.

which places secured financing on one side of the divide and contract law on the other seems entirely artificial. The distinction between market and nonmarket economics may be as arbitrary as it is uninteresting.

B

I want to try to convey some sense of the economic analysis of "nonmarket" law. Its basic premises are two:

1) People act as rational maximizers of their satisfactions in making such nonmarket decisions as whether to marry or divorce, commit or refrain from committing crimes, make an arrest, litigate or settle a lawsuit, drive a car carefully or carelessly, pollute (a nonmarket activity because pollution is not traded in the market), refuse to associate with people of a different race, fix a mandatory retirement age for employees.

2) Rules of law operate to impose prices on (sometimes subsidize) these nonmarket activities, thereby altering the amount or character of the activity.

A third premise, discussed at greater length later, guides some research in the economics of nonmarket law:

3) Common law (i.e., judge-made) rules are often best explained as efforts, whether or not conscious, to bring about either Pareto or Kaldor-Hicks efficient outcomes.

The first two premises lead to such predictions as that an increase in a court's trial queue will lead to a reduction (other things being equal—a qualification applicable to all my examples) in the number of cases tried, that awarding prejudgment interest to a prevailing plaintiff will reduce settlement rates, that "no-fault" divorce will redistribute wealth from women to men, that no-fault automobile accident compensation laws will increase the number of fatal accidents even if the laws are not applicable to such accidents, that substituting comparative for contributory negligence will raise liability and accident insurance premium rates but will not change the accident rate (except insofar as the increase in the price of liability insurance results in fewer drivers or less driving), that increasing the severity as well as certainty of criminal punishment will reduce the crime rate, that making the losing party in a lawsuit pay the winner's attorney's fees will *not* reduce the amount of litigation, that abolition of the reserve clause in baseball did not affect the mobility of baseball players (the Coase theorem, restated as a hypothesis), that the 1978 revision of the bankruptcy laws led to more personal-bankruptcy filings and higher interest rates, and that abolishing the laws that forbid the sale of babies for adoption would reduce rather than increase the full price of babies.

I have given a mixture of obvious and nonobvious hypotheses derived from my basic premises. Notice that I do not say intuitive and counterintuitive hypotheses, because all are counterintuitive to people who believe, as many economists and most lawyers do, that people are not rational maximizers except when transacting in explicit markets, or that legal rules do not have substantial incentive effects, perhaps because the rules are poorly communicated or the sanctions for violating them are infrequently or irregularly imposed.

C

Thus far in my discussion of the economic analysis of legal regulation of nonmarket behavior I have focused on the effects of legal change on behavior. One can reverse the sequence and ask how changes in behavior affect law. To make this reversal, though, one needs a theory of law, parallel to the rational-maximization theory of behavior. The economic theory of the common law, defined broadly as law made by judges rather than by legislatures or constitutional conventions or other nonjudicial bodies, is that the common law is best understood not merely as a pricing mechanism but as a pricing mechanism designed to bring about an efficient allocation of resources, in the Kaldor-Hicks sense of efficiency.[7] This theory implies that when behavior changes, law will

[7] See my book (pt. 2); and William Landes and myself (1987).

change. Suppose that at first people live in very close proximity to each other. Natural light will be a scarce commodity in these circumstances, so its value in exchange may well exceed the cost of enforcing a property right in it. Later, people spread out, so that the value of natural light (in the economic sense of value—exchange value rather than use value) falls; then the net social value of the property right (i.e., the value of the right minus the cost of enforcing it) may be negative. These two states of the world correspond roughly to the situations in England and America in the eighteenth century. The English recognized a limited right to natural light; they called this right "ancient lights." When American courts after independence decided which parts of the English common law to adopt, they rejected the doctrine of ancient lights—as the economic theory of the common law predicts they would.

Another example is the adoption of the appropriation system of water rights in the arid American West. In wet England and the wet eastern United States, the riparian system prevailed. This was a system of communal rights, which is a kind of halfway house between individual rights and no rights, and is inefficient for scarce goods. The appropriation system is one of individual rights, and was and is more efficient for areas that are dry (i.e., where water is scarce rather than plentiful)—which is where we find the appropriation system, as the economic theory of common law predicts. Or consider the different responses of the eastern and the western states to the problem of fencing out vs. fencing in. Fencing out refers to a property rights system in which damage caused by straying cattle is actionable at law only if the owner of the crops or other goods damaged by the cattle has made reasonable efforts to fence. Fencing in refers to a system where this duty is not imposed, so that the owner of the cattle must fence them in if he wants to avoid liability. The former system is more efficient if the ratio of crops to cattle is low, for then it is cheaper for the farmer than the rancher to fence. If the ratio is reversed, fencing in is a more efficient system. In fact, the cattle states tended to adopt fencing out, and England and the eastern states fencing in. Many similar examples could be given.[8]

Two objections to this branch of economic analysis of law must be considered:

1) One is that a theory of law is not testable, because when one is examining the effects of behavior on law rather than of law on behavior, the dependent variable tends not to be quantitative: it is not a price or output figure but a pattern of rules. However, the scientific study of social rules is not impossible; what else is linguistics? Fencing in vs. fencing out (or ancient lights vs. no ancient lights, or riparian vs. appropriative water rights) is a dichotomous dependent variable, which modern methods of statistical analysis can handle. And if a continuous variable is desired, it can be created by using the year in which the particular law was adopted (earlier adoption implying a more strongly supported law), the severity of the sanctions, or the expenditures on enforcement, to distribute states or nations along a continuum.

2) James Buchanan (1974), along with a number of neo-Austrian economists, holds that law should not be an instrumental variable designed to maximize wealth. Judges should not be entrusted with economic decisions—they lack the training and information to make them wisely. They should use custom and precedent to construct a stable but distinctly background framework for market and nonmarket behavior. But this is an objection to normative economic analysis of law—to urging, for example, that the common law (and perhaps other law) be changed to make it approximate the economic model of efficient law better—and the more interesting and promising aspect of economic analysis of law is the positive. I say this not because of a general preference for positive to normative inquiry, but because so little of a systematic nature is known about law. Law is not so well understood that one can hold a confident opinion about whether the right way to improve it is to

[8] See sources cited in fn. 7, from which the above examples are taken.

make the judges more sophisticated economically or more obedient to precedent and tradition.

II

Much of what I have said so far is old hat, at least to those familiar with the law and economics movement, so let me turn to some novel applications of economic analysis to law: applications to free speech and religious freedom, respectively.

A

It has long been recognized that the process by which truth emerges from a welter of competing ideas resembles competition in a market for ordinary goods and services: hence the influential metaphor of the "marketplace of ideas." It is also well known that because of the incompleteness of patent and copyright law as a system of property rights in ideas, the production of ideas frequently generates external benefits. Aaron Director (1964) and Ronald Coase (1974) have emphasized the peculiarity of the modern "liberal" preference for freedom in the market for ideas to freedom in markets for ordinary goods and services (both freedoms having been part of the nineteenth-century concept of liberty), and have attributed this preference to the self-interest of intellectuals.

Economists have paid scant attention, however, to the details of legal regulation in this area. Over the past seventy years or so, the courts have developed an elaborate body of doctrine through interpretation of the First Amendment's guarantee of free speech. Both the effects of this body of doctrine on the marketplace of ideas and the economic logic (if any) of the doctrines present interesting issues for economic analysis.

So far as effects are concerned, I suspect they have been few. Despite the high-flown rhetoric in which our courts discuss the right of free speech, they have countenanced a large number of restrictions—on picketing, on obscenity, on employer speech in collective bargaining representation elections, on commercial advertising, on threats, on defamatory matter, and on materials broadcast on radio and television. Although Americans appear to enjoy greater freedom of speech than citizens of the Western European nations, Japan, and other democratic nations at an equivalent level of development to the United States, the gap appears to have narrowed, not broadened, since the Supreme Court began to take an aggressive stance toward protection of free speech in the 1940's. It may be that as nations become wealthier and their people better educated and more leisured, the gains from restricting free speech—gains that have to do mainly with preserving social and political stability —decline relative to the costs in hampering further progress and in reducing the welfare of producers and consumers of ideas. These trends, I conjecture, are sufficiently pronounced to bring about (save possibly in totalitarian counties) dramatic increases in free speech regardless of the specifics of free-speech law.

The American law[9] has several interesting economic characteristics.

1) In the evolution of free-speech law, the first mode of regulation to go is censorship of books and other reading matter; the law's greater antagonism to censorship than to criminal punishment or other *ex post* regulation (for example, suits for defamation) being expressed in the rule that "prior restraints" on speech are specially disfavored. Censorship is a form of *ex ante* regulation, like a speed limit. The less common the substantive evil (the costs resulting from an accident due to carelessness, in the case of the speed limit, or the costs resulting from a treasonable or defamatory newspaper article, in the case of censorship), and also the more solvent the potential injurer,[10] the weaker the case for *ex ante* regulation is. With the growth of education and political stability, the social dangers of free speech have de-

[9] Well summarized, and in a form accessible to nonlawyers, in Geoffrey Stone et al. (1986, pt. 7).

[10] If the probability of apprehension and punishment is substantially less than one, the expected punishment may not deter wrongdoing even if the punishment, when imposed, takes away the offender's entire wealth and utility.

clined; and suppose the fraction of books and magazine articles that contain seriously harmful matter is today very small. Then the costs of a scheme in which a publisher must obtain a license from the public censor to publish each book are likely to swamp the benefits in weeding out the occasional prohibitable idea, especially since publishers have sufficient resources to pay fines or damage judgments for any injuries they inflict. It makes more economic sense in these circumstances to rely on *ex post* regulation (through criminal punishment or tort suits) of those ideas that turn out to be punishable. Censorship is retained, however, in areas, such as that of classified government documents, where the probability of harm is high and where in addition the magnitude of the harm if it occurs may be so great (for example, from disclosing sensitive military secrets) that the threat of punishment will not deter adequately because the wrongdoer will lack sufficient resources.

Many of these arguments could of course be made against *ex ante* regulation of safety, as by the Food and Drug Administration and OSHA. One difference is that while the First Amendment forbids overregulating the marketplace of ideas (and also, as we are about to see, the religious marketplace), no constitutional provision seems directed at forbidding overregulation of markets in conventional goods and services.

2) Consider now the onerous limitations that the Supreme Court has placed on efforts to sue the media for defamation. If we assume that news confers external benefits, then, since a newspaper or television station cannot obtain a significant property right in news, there is an argument for subsidizing the production of news. A direct subsidy, however, would involve political risks—though we have run them occasionally, as in the establishment of the Corporation for Public Broadcasting. A form of indirect subsidy is to make the victims of defamation bear some of the costs of defamation that the tort system would otherwise shift to the defamer. Notice, however, the curious effect of this method of subsidization, which may make it on balance inefficient. Because it is impossible to insure one's reputation, the victims of defamation cannot spread the costs of being defamed to other members of the community. The costs are concentrated on a narrow group, resulting in a deadweight loss if risk aversion is assumed. Moreover, public service is made less desirable, resulting in a decline in the quality of government. It would be difficult to prevent the decline by raising government salaries. The salary increase would have to be large enough to cover not only the expected cost of uncompensated defamation, but also the risk premium that risk-averse people would demand because they cannot buy insurance. Even if salaries are raised, the composition of public service will shift in favor of risk preferrers and people with little reputation capital. Finally, the difficulty of monitoring government outputs leads to heavy emphasis on economizing on visible inputs, for example by paying low salaries to government officials; and the problem of false economies is aggravated if the costs of government service are raised by curtailing the right of government officials to protect their reputations through suits for defamation.

3) The Supreme Court has distinguished between public and private figures, giving private figures a broader right to sue for defamation than public ones. This distinction may make economic sense. The external benefits of information about public figures are greater than those of information about private figures, and therefore the argument for allowing some of the costs to be externalized is stronger. Moreover, a public figure, being by definition newsworthy, has some substitute for legal action: he can tell his side of the story, which the news media will pick up.

4) A related point is that if the main reason for limiting efforts by government to regulate the marketplace of ideas is to foster the provision of external benefits, we would expect, and to a certain extent find, that the limitations on regulation are more severe the greater the likelihood of such benefits. Consider: Maximum protection for freedom of speech is provided to scientific and political thought, in which property rights cannot be obtained. Slightly less protection is given art, which enjoys a limited property right under

the copyright laws.[11] Even less constitutional protection is given to pornography and commercial advertising. And none is given to threats and other utterances that manifestly create net external costs.

Pornography appears to create no external benefits (no one but the viewer or reader himself benefits—and he pays), and may create external costs. Commercial advertising, a particularly interesting case, also creates few external benefits—since most such advertising is brand-specific and its benefits are captured in higher sales of the advertised brand—and it creates some external costs: competitor *A*'s advertising may go largely to offset *B*'s, and vice versa. This analysis implies that if the logic of free-speech law is basically an economic logic, commercial advertising that is not brand-specific, such as advertising extolling the value of prunes as a laxative, would receive greater legal protection than brand-specific advertising.

B

The First Amendment also forbids the government to make any law (1) respecting an establishment of religion or (2) prohibiting the free exercise of religion. The Supreme Court has enforced both clauses aggressively in recent years.[12] The economic effects of the Court's doctrines as well as their possible economic logic are interesting topics that economists (with the partial exception of Adam Smith) have not addressed.

There is, it is true, a nascent economic analysis of religion. Corry Azzi and Ronald Ehrenberg (1975) have formulated a simple (maybe too simple, given the variety of religious beliefs) economic model of religion, which assumes that people want to increase their expected utility from a happy afterlife.[13] The model leads to such predictions as that women will spend more time in church than men because the cost to women in foregone earnings is less, and that men will spend more time in church as they get older because as they approach the end of their working life it is optimal for them to switch from investing further in their earning capacity to investing in the production of afterlife utility. The authors find support in the data for their predictions.[14] My focus is different. I ask, what have been the effects on religious belief and observance of the Supreme Court's enforcement of the First Amendment? To avoid potential misunderstanding, I emphasize that I am offering no opinion on either the validity of any religious belief or the legal soundness of any of the Court's decisions.

Three major strands in the Court's modern decisions should be distinguished:

1) In its school-prayer decisions, and other decisions under the establishment clause, the Court has interpreted the concept of an "establishment" of religion very broadly, in effect forbidding the states and the federal government to provide direct support, financial or even symbolic, for religion. These decisions make a kind of economic sense, though perhaps only superficially. Public education (the principal arena of modern disputes over establishment of religion) involves the subsidizing of schoolchildren and their parents. Parents willing to pay the full costs of their children's education can and often do send their children to private schools. If they choose a public school instead, this may be because some of the costs will be paid by others, including persons who do not have school-age children as well as taxpayers in other parts of the state or nation. The principal economic argument for externalizing some of the costs of education is that education (with possible exceptions, as for vocational education and "phys. ed.") confers external benefits; that we all (or most of us, anyway) benefit from living in a nation whose popu-

[11] Only the specific work of art is protected; an artistic innovation (perspective, chiaroscuro, the sonnet, blank verse, etc.) is not.

[12] See Stone et al. (pt. 8).

[13] See also Ehrenberg (1977); Paul Pautler (1977); Barbara Redman (1980).

[14] For criticism of some of their results, see Holley Ulbrich and Myles Wallace (1984).

lation is educated. Therefore, to justify on economic grounds a public school's spending money on prayer and other religious activities, either these activities would have to be shown to produce positive externalities also (as by making schoolchildren more moral, or at least better behaved in school), or there would have to be economies from combining secular and religious instruction in the same facility, or private persons would have to volunteer to pay the incremental cost of the public school's religious activities, so that there would not be a subsidy.

If the Supreme Court were willing to accept any of these justifications—provided, of course, that they were adequately supported by evidence—then one might conclude that the Court was taking an economic approach to the issue in religion in the public schools. But, in fact, the modern Court forbids virtually every public school religious activity, whether or not any of these justifications is present. If none is present, it can indeed be argued that religious persons would be enjoying a public subsidy of religion if the activity were permitted. Parents willing to pay the full costs of education in a school that conducts prayer or engages in other religious activities can always send their children to a private school that offers such activities, thereby bearing the full cost of those activities rather than shifting a part of it to others in the community. Concern with public subsidies of religion may explain the Court's insistence that Christmas nativity scenes supported by public funds have a secular purpose, that is, confer benefits on nonreligious as well as religious persons. But the Court has not worried about the fact that the benefits may be greater for the latter persons, so that an element of subsidy remains. Nor has it explained its unwillingness to search for similar secular justifications for public school religious activity—such justifications as reducing the rowdiness of schoolchildren.

Further complicating the picture, the Supreme Court has declined to hold that the exemption of church property from state and local taxes is an unconstitutional establishment of religion. However, the consequence of the exemption is that the churches receive public services for which they do not pay. This is fine if they generate benefits for which they cannot charge, but the Court has not required that they show that. So here may be a large judicially sanctioned public subsidy of religion.

2) In its "free exercise" decisions, the Court has sometimes required public bodies to make costly accommodations to religious observance. An example is forbidding the denial of unemployment benefits to a person whose religion forbids him to accept a job offer that would require working on Saturdays. So the Court with one hand (establishment clause cases) forbids the subsidizing of religion and with the other (free-exercise cases) requires such subsidies.

3) In cases involving contraception, abortion, illegitimacy, obscenity, and other moral questions about which religious people tend to hold strong views, the Court in recent years has almost always sided with the secular against the religious point of view.

The decisions in both groups 1 and 2 favor religious rivalry or diversity (not competition in the economic sense: as we shall see in a moment, to subsidize rivalry as in 2 retards rather than promotes competition in the economic sense). Any public establishment of religion will tend to favor major religious groups over minor ones and can thus be compared to government's placing its thumb on the scales in a conventional marketplace, by granting subsidies or other benefits to politically influential firms. Refusing to accommodate fringe religious groups will have effects similar to those of establishing a religion because employment policies, and other public policies and customs, are chosen to minimize conflict with the dominant religious groupings.[15] It is no accident that the official day of rest in this country is the sabbath recognized by the mainline Christian groups. Fringe groups will therefore benefit from a rule requiring accommodation of their needs.

[15] As stressed in Michael McConnell (1985).

But since the costs of accommodation are borne by employers, consumers, taxpayers, other employees, etc., the group 2 cases actually subsidize fringe religious groups. And since it is no more efficient for government to subsidize weak competitors than strong ones, it may not be possible to defend the accommodation cases by reference to notions of efficiency. In addition, the group 1 cases may go further than necessary to prevent public subsidies of established religious groups, by neglecting the various justifications that might be offered for public support of religion—although allowing the property-tax exemption may correct (or for that matter, overcorrect) that tendency. The most important point to note, however, is that the Supreme Court has required government to subsidize fringe religious groups both directly and by discouraging religious establishments that inevitably would favor the beliefs and practices of the dominant sects in the community. By doing these things, the Court probably has increased religious diversity and may therefore have promoted religion, on balance, notwithstanding the "antireligion" flavor of some of its establishment cases.

The group 3 decisions favor religion, too —more precisely, private religious organizations—but in a subtler sense, which may be entirely unintended, even unrecognized, by the courts. By marking a powerful agency of government (the federal judiciary) as secularist, and, more important, by undermining traditional values through invalidation of regulations that express or enforce those values, these decisions increase the demand for organized religion, viewed as a preserver of traditional values. If the government enforced the value system of Christianity, as it used to do, people would have less to gain from being Christian. The group 1 cases have a similar effect. By forbidding teachers paid by the state to inculcate religious values, the courts have increased the demand for the services provided by religious organizations. And allowing the property-tax exemption lowers the costs of these organizations.

Of course, there may be no net increase in the provision of religious services if a public school in which teachers lead prayers or read to students from the Bible is treated as a religious organization, but my concern is with the effect on private organizations. Similarly, a government that rigorously repressed abortion might be thought of as the enforcement arm of the Christian sects that regard abortion as immoral; but by thereby assuming one of the functions of private religious organizations, it would be competing with those organizations and thus reducing the demand for the services provided by them.

There is a further point. As Adam Smith pointed out (1937, pp. 740–50), the effectiveness of a private group's monitoring and regulating the behavior of its members is apt to be greater, the smaller the group (this is the essence of cartel theory), from which Smith inferred that the more religious sects there were, and hence the smaller each one was on average, the more effective would religion be in regulating behavior. This implies that legal regulations which have the effect of atomizing rather than concentrating religious organization may improve the society's moral tone even if they diminish the role of government in inculcating moral values directly.

It may be hard to believe that the moral tone of our society has actually improved since the Supreme Court adopted its aggressively secularist stance, but economic analysis suggests that the situation might be worse rather than better if the Court had weakened private religious organizations by allowing government to compete more effectively with them in inculcating or requiring moral behavior. Since government and organized religion are substitutes in promoting moral behavior, an expansion in the government's role as moral teacher might reduce the demand for the services of organized religion. I say "might" rather than "would" because, to the extent that the government's role as moral teacher is taken seriously, a government that seeks to promote religiously based moral values may help "sell" religious values, and the organizations that promote them, over their secular substitutes. But this assumes what history suggests is unlikely: that the government will find a way of supporting religion on a genuinely nonsectarian basis rather than establishing a particular sect and thereby weakening competing sects and maybe religion as a whole.

To prove, in the face of the conventional wisdom to the contrary, that the Supreme Court's apparently antireligious decisions have promoted religion would be a formidable undertaking, and here I offer only two fragments of evidence. The first is the rapid growth in recent years of evangelical Christianity, formerly a fringe religious grouping and one marked by emphatic adherence to traditional values.[16] The second is the startling difference in religiosity between the United States and Western Europe. Not only does a far higher percentage of Americans believe in an afterlife than the population of any western European country other than Ireland,[17] but this percentage has been relatively constant in the United States since the 1930's, while it has declined substantially in Europe over the same interval.[18] Almost all Western European nations have an established (i.e., a taxpayer-supported and legally privileged) church (or churches, as with the state churches of Germany), and some require prayer in public schools.[19] To the extent that establishment discourages the rise of rival sects, it reduces the religious "product variety" offered to the population, and I would expect the demand for religion to be less. The American system fosters a wide variety of religious sects. Almost every person can find a package of beliefs and observances that fits his economic and psychological circumstances. And by preventing the government from playing a shaping role in the moral sphere the Supreme Court in recent years has, I have conjectured, increased the demand for religion as a substitute institution for the regulation of morals.

No doubt the Supreme Court's causal role in all this is smaller than I have suggested. The tradition of religious diversity in the United States is very old, and the Court's contribution to maintaining it may be slight. Nevertheless, economic analysis suggests that the religious leaders who denounce the course of the Court's decisions and the secular leaders who defend it may be arguing contrary to their institutional self-interest.

C

My discussions of free speech and religion can be connected as follows. One possible reading of the First Amendment (I do not suggest the only, or a complete one) is that it forbids government to interfere with the free market in two particular "goods"—ideas, and religion. Government may not regulate these markets beyond what is necessary to correct externalities and other impediments to the efficient allocation of resources. This seems an appropriate description of how modern courts interpret the amendment; the principal though not only exceptions are the cases that forbid what might be called "efficient" establishments (establishments that do not involve a subsidy to religious persons beyond what can be justified on secular grounds) and the cases requiring accommodation of religion in the sense of subsidizing fringe religious groups. There is no compelling economic argument for such a subsidy unless something can be made of Adam Smith's point that the more separate religious sects there are, the more effective religion is in bringing about moral behavior—and morals supplement law in correcting negative externalities such as crime and fostering positive ones such as charity.

But a lecture is not the place to prove a new economic theory. All that is feasible is to suggest that a particular theory holds promise and is thus worth pursuing. I hope I have persuaded you that what may loosely be called the economic theory of law has a significant potential to alter received notions, generate testable hypotheses about a variety of important social phenomena, and in short enlarge our knowledge of the world.

[16] See *The Gallup Report* (1985, pp. 3, 11).

[17] See *The Gallup Report*, p. 53.

[18] See *The Gallup Report*, pp. 9–10, 40, 42, 53.

[19] On the religious establishments of Western Europe, see, for example, E. Jürgen Moltman (1986); E. Garth Moore (1967); Franklin Scott (1977, pp. 571–75); Frederic Spotts (1973).

REFERENCES

Ashenfelter, Orley and Rees, Albert, *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973.

Azzi, Corry and Ehrenberg, Ronald, "Household Allocation of Time and Church Attendance," *Journal of Political Economy*, February 1975, *83*, 27–56.

Becker, Gary S., *The Economic Approach to Human Behavior*, Chicago: University of Chicago Press, 1976.

______, *Human Capital: A Theoretical and Empirical Analysis, With Specific Reference to Education*, NBER, New York: Columbia University Press, 2d ed., 1975.

______, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

Buchanan, James M., "Good Economics—Bad Law," *Virginia Law Review*, March 1974, *60*, 483–92.

Coase, Ronald H., "Economics and Contiguous Disciplines," *Journal of Legal Studies*, June 1978, *7*, 201–11.

______, "The Market for Goods and the Market for Ideas," *American Economic Review Proceedings*, May 1974, *64*, 384–91.

Director, Aaron, "The Parity of the Economic Market Place," *Journal of Law and Economics*, October 1964, *7*, 1–10.

Ehrenberg, Ronald G., "Household Allocation of Time and Religiosity: Replication and Extension," *Journal of Political Economy*, April 1977, *85*, 415–23.

Ehrlich, Isaac, "Participation in Illegitimate Activities: An Economic Analysis," in Gary S. Becker and William M. Landes, eds., *Essays in the Economics of Crime and Punishment*, NBER, New York: Columbia University Press, 1974, 68–134.

Ellis, John M., *The Theory of Literary Criticism: A Logical Analysis*, Berkeley: University of California Press, 1974.

Fogel, Robert W. and Engerman, Stanley L., *The Reinterpretation of American History*, New York: Harper & Row, 1971.

Hirshleifer, Jack, "The Expanding Domain of Economics," *American Economic Review*, December 1985, Suppl., *75*, 53–68.

Landes, William M. and Posner, Richard A., *The Economic Structure of Tort Law*, Cambridge: Harvard University Press, forthcoming 1987.

McConnell, Michael, "Accommodation of Religion," *Supreme Court Review*, Chicago: University of Chicago Press, 1985, 1–59.

Moltmann, E. Jürgen, "Religion and State in Germany; West and East," *Annals of the American Academy of Political and Social Science*, January 1986, *483*, 110–17.

Moore, E. Garth, *An Introduction to English Canon Law*, Oxford: Clarendon Press, 1967.

Pautler, Paul A., "Religion and Relative Prices," *Atlantic Economic Journal*, March 1977, *5*, 69–73.

Posner, Richard A., *Economic Analysis of Law*, Boston: Little, Brown, 3d ed., 1986.

Pyle, David J., *The Economics of Crime and Law Enforcement*, New York: St. Martin's Press, 1983.

Radnitzky, Gerard and Bernholz, Peter, *Economic Imperialism: The Economic Approach Applied Outside the Field of Economics*, New York: Paragon House, 1986.

Redman, Barbara J., "An Economic Analysis of Religious Choice," *Review of Religious Research*, Summer 1980, *21*, 330–42.

Scott, Franklin D., *Sweden: The Nation's History*, Minneapolis: University of Minnesota Press, 1977.

Scott, Robert E., "A Relational Theory of Secured Financing," *Columbia Law Review*, June 1986, *86*, 901–77.

Smith, Adam, in Edwin Cannan, ed., *The Wealth of Nations*, London: Methuen, 1937.

Spotts, Frederic, *The Churches and Politics in Germany*, Middletown: Wesleyan University Press, 1973.

Stigler, George J., "Economics—The Imperial Science?," *Scandinavian Journal of Economics*, No. 3, 1984, *86*, 301–13.

______, "The Theory of Economic Regulation," *Bell Journal of Economics*, Spring 1971, *2*, 3–21.

Stone, Geoffrey R. et al., *Constitutional Law*, Boston: Little, Brown, 1986, pts. 7 and 8.

Ulbrich, Holley and Wallace, Myles, "Women's Work Force Status and Church Attendance," *Journal for the Scientific Study of Religion*, December 1984, *23*, 341–50.

The Gallup Report, Report No. 236, Princeton: The Gallup Poll, May 1985.

ROUNDTABLE ON TEACHING UNDERGRADUATE COURSES IN QUANTITATIVE METHODS

The Centrality of Economics in Teaching Economic Statistics

By GLEN G. CAIN*

The elementary course in economic statistics must devote more time to statistics than to economics. Students who major in economics will take at least six courses in economics, but the course in economic statistics may be their only course in statistics. Nevertheless, the basic purposes of economics should be kept in mind in shaping the course. Although the statistical methods used in applied economics deal with measurable variables, it is my view that the ultimate purpose of economics as an applied social science is the improvement of human welfare, which is intrinsically an unmeasurable concept.

Our recognition of the unattainable ideal does not stop us from obtaining less than ideal measurements that are nevertheless useful. The two basic problems in this task of economic measurement are (a) to define and measure the correct outcomes, and (b) to measure their determinants so that we can predict and, ideally, influence the outcomes. The contribution that the science of statistics can make to these tasks is fundamentally that of the art and craft of using observed sample data to make inferences about unknown population parameters in economics. Inferential statistics is fundamental. Descriptive statistics, which is the art and craft of organizing and summarizing sample data, is useful and necessary for learning inferential statistics, but it is not fundamental in its own terms.

Turning to the structure of the course in economic statistics, let us apply a principle of economics to its teaching by specifying the constraints under which we seek to optimize our goals. Four major constraints face the teacher of elementary economic statistics: 1) the limited time in a one-semester course; 2) the typically large size and heterogeneity of the class; 3) the limited economic and, especially, mathematical background of most of the students; and 4) the limited skills of the teacher.

A word about constraint 4, which implicitly qualifies much of what follows. Instructors of economic statistics have varying talents for and preferences about the course, and it is appropriate to play to one's strengths. If someone is a whiz at teaching probability, this topic by this instructor may captivate students and inspire them toward an understanding of inferential statistics. Another instructor may be skillful in using examples from such economic topics as income distribution or macroeconomics to illustrate how economists use sample evidence to estimate and test interesting relationships between outcome variables and their determinants. We each have our own styles of teaching, and my suggestions for content and methods should be viewed as subservient to any particular instructor's tastes and skills. With that qualification, I turn next to the content of the course.

I. Course Content in Terms of Course Objectives

Four objectives and their implications for course content are listed and discussed below.

1) Students should obtain some degree of economic literacy or, to use a newly coined phrase, "economic numeracy," about the principal indicators of the performance of an economy. Both professional economists and the popular media place great emphasis on

*University of Wisconsin, Madison, WI 53706. I am grateful to Gary Chamberlain, Chris Flinn, Lee Hansen, and John Rust for helpful comments.

statistical indicators such as GNP, per capita income, poverty measures, the Consumer Price Index (CPI), inflation rates, unemployment rates and other labor force measures, and trends and other measures of change in all these variables. Each indicator involves statistical definitions and properties that are informative of their strengths and weaknesses, and that are unlikely to be satisfactorily explained to students in other courses.

Time constraints prohibit comprehensive explanations of all these indicators, and I recognize that many instructors will not give this objective a high priority. However, I believe that analyzing a few of these indicators can be useful in several ways. For example, the CPI involves the statistical techniques of summations, weighted averages, and survey sampling. The CPI also invites a discussion of its economic purposes; in particular, its use as a measure of the elusive concept of the cost of living. The inherent ambiguity of the CPI as such a measure can be shown, and students can gain an understanding of the famous "index number problem" in a practical context.

The rate of unemployment is another important indicator. What are the purposes of the unemployment rate, and how well does it serve these purposes? In answering these questions we need to know how unemployment is measured. The distinctions between reliability and sampling error on the one hand and validity and nonsampling errors on the other hand are demonstrable, since the unemployment rate is so convincingly reliable for the nation as a whole (although not for local areas), while its validity is highly controversial. When students understand the operational definition of unemployment, they will realize that the official measure is deficient in many ways. But they may also come to view the official unemployment rate as an ingenious and valuable measure of the performance of the economy.

2) A second and more important objective of an elementary course in economic statistics is to introduce students to how economists use statistical methods in their research. Descriptive statistics play an essential role in presenting the economic indicators mentioned above. As noted, the CPI is one type of average. The unemployment rate is another; namely, the mean of a qualitative or dummy variable. Incidentally, dummy variables are especially useful to develop early in the introductory course, partly because they serve to disabuse the students of the misconception that economic statistics deals only with variables that are conventionally quantifiable. The statistics of income distribution serve to introduce all sorts of measures of dispersion, and so on with other economic statistics.

Most research economists are more interested in measuring relationships between and among variables, than in describing a single variable. There is much value in using descriptive statistical techniques, such as regression analysis, to describe relations among variables. Certainly, the trend line for a time-series illustrates the use of simple regression that is easily understood by beginning students. If the instructor has more courage, or should I say, a stronger stomach, the "Phillips Curve" showing the time-series relation between inflation and unemployment can illustrate a linear or nonlinear regression function.

As noted above, the task of making inferences about these relationships is more fundamental. It is also much more difficult because it takes on the formidable problem of measuring causal relations between economic variables. If the word "causal" intimidates you, as it does me, substitute the term "stable-predictive," meaning that the predictive association of one variable to another is stable in relevant contexts. Indeed, I view the definition or specification of the relevant context as essential for understanding these issues of stability and causality. Is the context a cross section or a time-series? Do we wish to achieve stability in an "associational" relationship, where causality may not be of concern? Or do we seek to predict a response to an autonomous change in an independent variable, and, if so, what is the institutional context of such a change? Is the sample we use to estimate this latter relationship appropriate for an inference to the population in which this institutional context is embedded? These are difficult questions to pose, and I personally have difficulty seeing them answered in most of the economic research that I read.

Often, the mechanical operations in calculating a variety of relationships between variables—for example, a correlation—may be easy; with computers, exceedingly easy. What is difficult is to persuade oneself and others that the calculated relationship is causal, or if not causal, at least stable.

There are two well-known obstacles to establishing causality in economics. The first is that economic theories seldom tell us all the variables that determine an outcome of interest. The second is the inability of economists to conduct controlled experiments to measure relationships between, say, an "input" (or explanatory) variable and an outcome (or dependent) variable. In place of controlled experiments, economists, like sociologists and even such physical scientists as medical scientists working on problems of disease, must rely on observational studies.

An implication of the economist's reliance on observational studies is that many variables have to be taken into consideration when predicting or explaining an outcome. What is more important, and less obvious to students, is that even if we want to focus only on how a single variable, such as years of schooling, affects an outcome, such as income, we still need to allow for many variables. The natural world, which for economists is usually the market place, provides us with variation in schooling that is intertwined with variation in other determinants of income. To deal with this problem we need to introduce the multiple regression model or its close relatives in statistics.

Should multiple regression be covered in an elementary economics statistics course? Before defending this notion, I need to discuss the other two goals.

3) The third purpose is to introduce students to the basic concepts of elementary statistics, which provide the tools to achieve the first two goals and which apply generally to all fields of empirical science. This point returns us to my opening remark that the course will devote more time to statistics than to economics. Most texts in economics and business statistics are not sharply distinguishable in their coverage, organization, and sequence of topics from texts that are used in elementary courses in a statistics department.

On the positive side, this reflects the necessity for coverage of descriptive statistics, for the basic ideas of estimation and hypothesis testing in inferential statistics, and for the attention to probability as a bridge between the two branches of statistics. On the negative side, the usual text usually relegates regression analysis to the last chapters, along with, perhaps, separate attention to time-series analysis. My impression is that simple regression receives short shrift and multiple regression no coverage at all in the typical course in economic or elementary statistics.

In light of the four constraints noted above, this dedication to a conventional statistics-department coverage of material is understandable. However, it raises two problems. An obvious one is that objectives 1 and 2 above are neglected. A second problem is that the statistics department rather than the economics department may be the proper source for this type of course.

The alternative I suggest is to rearrange the sequence and, in recognition of the constraints, to drop or diminish some aspects of the conventional agenda in objective 3 and to include more materials related to objectives 1 and 2. More on this alternative is discussed in the last section.

4) The fourth objective in an elementary statistics course, whether in economics or not, is to introduce the student to computers and to their role in both the computations and the pedagogy of elementary statistics. Personal computers, with their interactive features and their compactness, are a marvelous aid in a statistics course. The computational advantages are obvious, particularly for regression analysis. Large and realistic data sets are permitted, logarithmic and other transformations are facilitated, along with many other useful manipulations. Finally, I would emphasize that multiple regression computations are feasible. Clearly, these savings of time, the reductions of frustrating computation errors, and freeing the student to concentrate on the conceptual ideas all help students overcome their difficulties in

juggling required mathematics, statistical concepts, the economic content of the material, and computational chores.

Two additional benefits of computers will be mentioned briefly. They permit simulations to help students understand probability concepts and, in particular, the central limit theorem regarding the normality of the distribution of sums and of the sample mean. Second, personal computers are so likely to be owned or used in jobs by the economics majors after they graduate that a continued use of statistics after the final exam is over is highly likely.

II. Brief Comments on How to Cover the Contents Discussed Above

The conventional economic statistics course, judging by most of the texts, appears to be aimed at objectives 3 and 4, given the recent supplementation of texts with materials promoting computer uses. Objective 1 does not appear to be well served, judging by the reliance on the mere use of economic variables to illustrate computational problems. This is hardly a substitute for attention to the purposes of the economic statistical indicators, to their operational definitions, and to their reliability and validity.

The second objective, introducing students to how economists do economic research, ought to be central, yet I admit that this objective is hard to achieve, and it may require a more advanced course in introductory econometrics. Even so, my preferences are to try; to try to get across the elusive concept of causality in economics and how this concept has different meanings in different contexts; to try to deal with the harsh reality of reliance on observational studies and reliance on economic theory to specify interpretable statistical functions. Fortunately or unfortunately, the phrase "interpretable statistical function," at its simplest practical level, is usually going to consist of a multiple regression model. However, even if multiple regression is not covered, simple regression can be introduced as a part of descriptive statistics very early in the course. (Two texts which do this are: *Statistics*, W. W. Norton, 1978, by D. Freedman, R. Pisani, and R. Purvis; and *Elementary Statistics*, Duxbury, 1984, by R. Johnson.)

What must be sacrificed to get across these objectives? Despite some pedagogic productivity gains from the use of computers, my experience is that some traditional content must be deleted or reduced. Indeed, I have never covered all four objectives to my satisfaction. I usually slight probability, but I emphasize frequency distributions and histograms, which convey the concept of representing relative frequencies by the areas of the histogram. Except for frequency distributions, and the mean, median, and variance, I skim over most descriptive statistics. The normal distribution receives almost all my attention in teaching about probability distributions, and I quickly discuss the conditions for assuming the normality of sample means to justify this emphasis. Old-timers may recognize the influence of the old text, *Statistics: A New Approach* (The Free Press, 1956) by W. A. Wallis and H. V. Roberts. I usually skip analysis of variance, although my emphasis on regression and on dummy variables provides an alternative to analysis of variance. I also skip *chi*-square tests and tests for the equality of variances. Finally, I deemphasize correlation even while emphasizing regression, and I emphasize point and interval estimation over hypothesis testing.

I close on the sober note that I am still struggling with my strategy for teaching the course. And I remind you of my earlier point that each teacher has to play according to his or her own strengths.

Teaching Statistical Methods to Undergraduate Economics Students

By WILLIAM E. BECKER*

Little published work has appeared on the teaching of statistics in economics. This lack of attention is surprising since John Seigfried and James Wilkinson (1982) found that nearly 85 percent of the 546 schools surveyed required at least one course in statistics for an undergraduate major in economics, and approximately 50 percent of the departments of economics offered their own course. After introductory economics, only intermediate macroeconomics and money and banking have higher enrollments than statistics. This paper focuses on the goals and objectives of the introductory economics statistics course. It also gives attention to the more advanced econometrics courses undergraduates may elect to take.[1] It emphasizes the quantitative skills all undergraduate majors in economics should possess irrespective of their career plans.

As pointed out by W. Lee Hansen (1986), it is much easier to talk about what should be covered in a course than to define the student competencies to be gained. The latter requires the definition of student activities while the former treats the student as a passive recipient. If students are to do more than regurgitate definitions, duplicate proofs, or perform repetitive computations, then what is expected from them must be specified. Whether an instructor "covers" something or whether students can actually do something with what is "covered" are two different issues. This paper focuses on what students should be expected to do, not on what should be covered in economics statistics courses.

I. Applying Statistical Methods to Economic Data

Eric Sowey defines econometrics as the study of "...applying statistical methods to economic data" (1983, p. 257).[2] An economics department's rationale for offering an introductory statistics course or more advanced econometrics courses, as opposed to having them offered by a department of statistics or mathematics, rests on an argument that there is something unique about applying statistical techniques to economic data.[3]

*Indiana University, Bloomington, IN 47405. I am indebted to Suzanne Becker, Rendigs Fels, W. Lee Hansen, Donald Harnett, Thomas Johnson, Peter Kennedy, Catherine Melfi, and Robert Sandy for helpful comments on earlier drafts of this article.

[1] According to Seigfried and Wilkinson, the undergraduate econometrics courses had the seventeenth highest enrollment among the 62 courses listed. They were taught by 46.5 percent of the 512 departments for which information was available, but only 5.9 percent of the entire sample of 546 departments required an econometrics course for undergraduate majors in economics.

[2] Sowey defines econometrics as "the discipline in which one studies theoretical and practical aspects of applying statistical methods to economic data for the purpose of testing economic theories (represented by carefully specified models) and of forecasting and controlling the future path of economic variables" (p. 257). The words "practical aspects of applying" are downplayed by Sowey; he presents a highbrow view of what econometricians should teach, with the teaching of applications left to others.

[3] Donald Waldman and I (1987) prepared a graphical interpretation of probit analysis where the dependent variable was unmeasurable utility and only one of two outcomes was observable for each sample individual's choice. A statistician suggested that we redo the example using a "continuous real-world measure" as the dependent variable. He suggested that we first fit a least squares regression to this continuous measure and then split the data to form the dichotomous dependent variable for the probit model. Students would gain insight in comparing the results. While a statistician untrained in economics might think that students will find such comparisons enlightening, the unmeasurable nature of utility and the modeling of discrete choices is lost and thus has little relevance in a statistics course offered by an economics department.

Yet the word "applying" gets lost in many discussions about teaching courses labeled "economics statistics" and "econometrics." It is the application of statistical measures and statistical inference in an economic analysis that must be emphasized in an economics department's offerings.

To learn to apply the tools of statistics to economic data, there is little justification for examples involving the drawing of balls from urns, coin and dice tricks, or examples from the natural sciences. Students need to be involved in working exercises, considering case studies and solving problems which reflect what economists do. This implies replacing the urns with a preselection pool from which individuals are hired; replacing the coins and dice with surveys in which individuals face multiple choice responses, and replacing examples from genetics with examples from quality control.

Analyzing economic issues and drawing conclusions based on economic data may be the best, if not the only, way to sharpen the ability of students to "emphasize the interpretation, the limitations and the significance for economics of statistical techniques," which according to Sowey (p. 282) is a goal of "nonspecialist courses" in econometrics. Jacques Drèze (1983), commenting on Sowey's suggestion that applications should take a back seat to theory when class time is short in "specialist courses," stated that practical illustrations and examples should be introduced all along in the teaching of economics statistics. Both basic and more advanced principles can be put to practical use through examples drawn from the academic journals and popular press. The short or mini case study approach introduced into the teaching of economics by Rendigs Fels (1974) can be easily and fruitfully modified for teaching economic statistics. Short case studies enable instructors to demonstrate a specific form of statistical analysis while giving students an opportunity to observe how concepts and theory are used to examine particular problems or issues.[4] When confronted with a similar problem, it seems more likely that students can make the transfer than if the specific form of analysis had been presented without the case application.

There are those who argue that even undergraduate students should be able to work with large, real-world data sets on which a myriad of analytical techniques can be performed. Some introductory textbooks include a large data set which is used in each chapter. The mini case study approach of Fels, however, suggests that before students are confronted with compound problems and large data sets, they should be exposed to real or simulated situations that involve a limited number of concepts and easily managed data sets.

In using large data sets, it is difficult if not impossible for beginning students to see how an array of observations and a specific method of analysis can illuminate a question or problem. For example, in a histogram of 1,000 observations, students may not be able to appreciate the influence of extreme values on different measures of central tendency; with only a few well-chosen observations, the effect of extreme values is made apparent. More important, use of one large data set to show multiple statistical techniques may give the impression that a good

[4] As an example of a short case, consider the following from myself and Donald Harnett:

> "You are the state commissioner of insurance. An irate consumer group wants you to initiate a costly review process that it hopes will result in the fining of an insurance company for deceptive advertising. The insurance company is advertising that the average processing time to settle a household damage claim fully is only 8 days. The consumer group claims that the majority of such claims are not settled in even 13 days. Thus, the group asks 'How can the insurance company be guilty of anything short of fraud?' What action will you take before responding?" [1987, p. 81]

This case calls attention to the difference between the mean and median, their relationship to skewness, and the need for measures of dispersion. In the case approach, statistical concepts are taught only if they can be used in a problem-solving situation. This means that the first question asked by the instructor in constructing a syllabus is not what topics should be taught, but rather what type of problems and questions should the students be able to analyze and answer. Once the problem and questions are specified the statistical concepts, methods, and techniques to teach will be apparent.

analysis involves using all the statistical techniques available without regard to their appropriateness. Students should come to realize that the problem or question dictates the data needed and the analytical techniques employed, and not the other way around. Fels demonstrated that many short cases can help students recognize this point. Starting with a limited number of statistical concepts and analytical techniques, students learn how to select the concepts and techniques appropriate for solving different classes of problems. Since the emphasis is on problem solving, students quickly realize that memorization of a lot of definitions, the ability to plug numbers into formulas, and agility with a hand calculator are not critical in applying statistical methods in economics. To foster this realization, instructors must ensure that their presentations and instructional materials aim at problem solving and do not reflect only a preconceived notion about what topics must be covered.

II. Statistical Analysis Requires Computer Software

To apply statistical methods to economic data, students need computer programs for handling computations. What they must do with a program (arithmetic, transformations, graphics, search procedures, etc.) and how sophisticated the program must be (both in terms of programming knowledge and statistical routines) depends on the course. But if econometrics is viewed as an area of applied statistics, then all undergraduate econometrics courses starting with the introductory course should use computer programs.

Surveys completed by textbook publishers and academics alike suggest that instructors are already using computer programs such as MINITAB, SAS, and SPSS in statistics courses. How these programs are used remains in doubt. For example, in the E. L. Rose, J. A. Machak, and W. A. Spivey (1986) survey of business programs, 99 percent of the responding schools said they covered simple regression in the introductory statistics course and 88.8 percent considered multiple regression. Yet, 48.7 percent of the responding schools stated that they do not use the computer for regression with cross-sectional data and 57.3 percent do not use it for regression with time-series data. About 17 percent do not allocate any class time to using computers in statistics, and 68.4 percent allocate only 1 to 4 hours during the term for this purpose.

The first course in statistics should require students to load and run a menu-driven statistics package (for example, MICROSTAT on a CPM or DOS-based microcomputer, IDA on a VAX minicomputer, or STATPAK on an IBM mainframe). Menu-driven programs are advantageous because they do not require extensive class time for students to learn to operate the program. For example, a *PC Magazine* review of MICROSTAT stated that "someone completely unfamiliar with MICROSTAT could follow the clear menus and obtain useful output" (Marvin Bryan, 1986, p. 214). Unlike the detailed instructions required to run a regression on command-driven programs, menu-driven or interactive programs prompt the user for specification details (for example: What is the dependent variable? How many explanatory variables will be used? Identify the explanatory variables? etc.). Of course, to follow these menus, students must understand what they are trying to do, and once the output is obtained, they must be able to explain its meaning. Only then is the output useful. Unless students attempt statistical work on a computer, however, they will never appreciate the potential of statistical methods in economic analysis.

By the end of the first course in statistics, students should be able to use a menu-driven statistics package to enter and retrieve data, perform data transformations, calculate descriptive statistics, generate frequency distributions, determine probabilities with the binomial, hypergeometric, poisson, normal, *t*, *F*, and *Chi*-square distributions, and fit least squares regressions. Only a few hours of class time are typically required to help students gain the computer and program expertise required to perform these functions.

Time devoted to computer training also can be used to teach concepts which previously required student knowledge beyond high school algebra. For example, during the computer training, students can learn how the computer generates distribution functions without first learning how to integrate density functions. Students can learn how the regression algorithm is based on the idea of minimizing the sum of squared errors. Without ever taking a derivative, they can develop an understanding of least squares through discussions of computer-generated scatterplots.

Class time need not and should not be devoted to teaching computer use outside the context of statistics. Just as students should learn that the problem or issue under consideration gives rise to the data and statistical techniques employed, they should come to realize that the data and statistical techniques suggest the computer program to be used and not the other way around. This later student recognition may not be achievable in the first statistics course, but is achievable by the end of the more advanced econometrics course.

Unlike microcomputer software, use of statistical packages on minicomputers and mainframes has disadvantages for beginning students. Instructors must invest time to make arrangements to obtain and maintain student accounts, which typically expire after the course is over, prohibiting continued use by students. Moreover, unless future employers possess these programs and hardware, students will be unable to demonstrate their practical knowledge. Microcomputer-based programs overcome these disadvantages. Through the use of site licenses, student versions, education discounts, or public domain software, students can purchase copies of statistical packages at little more than the cost of a disk. Bundled textbook, workbook, and statistical packages are now selling for under $55.

The ease of menu-driven programs in generating descriptive statistics, calculating probabilities from specific distributions, and performing other routine computations assures their future use by students who continue their education in econometrics and those who terminate their training after the first course in statistics. Continuing students will need to have their library of computer programs expanded. Powerful and yet relatively easy-to-use programs, such as LIMDEP, RATS, and GAUSS, make it unnecessary for these students to learn structured programming languages to do matrix manipulations, perform maximum-likelihood estimation, test linear restrictions, or design their own statistics.

III. Outdated Computational Procedures

The time required to help students learn how to use menu-driven statistics programs can be more than offset by savings in instructional time devoted to teaching computational algorithms. With programs to determine probabilities based on appropriate distributions, beginning students no longer need the ability to approximate the binomial with the normal, use the binomial to approximate the hypergeometric or use the z when the t distribution is appropriate. Interpolating between values in z or t tables is a skill that students no longer need.

Students should no longer be expected to demonstrate their proficiency at plugging numbers into computational formulas. Instead, they should be called upon to demonstrate the principles involved in a computational procedure. Such learning does not imply that students know all the alternative computational versions of a formula. For example, students should know that the variance is the average squared deviation of observations around their mean. This definition of the variance implies that students calculate a finite population variance from the formula[5] $\sigma^2 = \Sigma(f_i/N)(x_i - \mu)^2$. To understand the definition of the variance,

[5] This is not to argue that the calculation of sample variances should be deemphasizèd, or that the calculation of the variance of a continuous random variable has no place in the introductory course. Until students understand degrees of freedom, however, calculations of sample variances need not lead to student understanding that the variance is nothing more than the average of the squared deviations of observations around their mean. Similarly, beginning students have difficulty grasping the meaning of the variance if they are first confronted with an integral which involves the product of a density function and the squared deviations of the continuous random variable around its mean.

however, students need not demonstrate proficiency with all the algebraically equivalent forms of this equation.

At more advanced levels of econometrics, computer programs make matrix manipulation relatively easy; this implies no need to emphasize hand calculator procedures. For example, in the estimation of regression coefficients, students no longer need to see or work with the transformation matrix

$$A=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}-\frac{1}{n}\begin{bmatrix}1 & 1 & \dots & 1\\ 1 & 1 & \dots & 1\\ \cdots & \cdots & \cdots & \cdots\\ 1 & 1 & \dots & 1\end{bmatrix}$$

which is used to convert the n observations on each variable (as contained in the X matrix and y vector) to deviations from sample means. A computational procedure that first requires converting data to deviations via this A matrix and then calculating slope coefficients in a second step and the intercept in a third is of no value (as a learning experience or for practical purposes) to students who have access to computer programs for matrix operations. The A matrix has nothing to do with understanding regression coefficient estimation, coefficient interpretation, and the role of coefficients in prediction. This understanding comes from working directly with the products of the X matrix, the $(X'X)^{-1}$ matrix and the $X'y$ vector. Diagrams showing how these matrices project observations into predictions and errors provide the intuition for least squares regression analysis.

Elimination of coefficient estimation based on the A matrix would remove 3 pages from the 39 pages in Chapter 5 of J. Johnston's (1984) econometrics text. Similar savings would be realized in other books where the k-variable linear model is introduced in matrix form. Eliminating discussions of all such outdated computational procedures frees class time to help students gain a better grasp of when and why particular statistical techniques may be appropriate, and makes more time available for the instructor to help students formulate correct interpretations of their results. It also provides class time to introduce the more contemporary modeling and estimation techniques associated with problems involving time-series, discrete choice, and qualitative data.

Modern computer statistical packages make it possible for students to use certain maximum-likelihood techniques which a few years ago could be performed and understood only by Ph.D. econometricians. Using computer printouts that show the iterative values tried by the program, students with an understanding of least squares regression can learn probit, logit, and Tobit modeling without first taking four semesters of calculus. For example, using a probit specification, Waldman and I demonstrate with a three-dimensional diagram the maximum-likelihood method. We show the line that best describes an unobservable but normally distributed variable y^* given the relative frequencies of observations that are known to be above (or below) a threshold value of y^*, at each value of an independent variable. This development and interpretation of the line's properties require an understanding of density, mass, and the maximizing principle, but not the actual calculation of integrals or derivatives.

In conclusion, the availability of easy-to-use and powerful microcomputer programs puts sophisticated statistical techniques within the reach of undergraduates. Practical applications should never again take a back seat to theory in the introductory statistics course or the more advanced econometrics courses. No longer should students walk away from these courses saying that econometrics is "a marvelous array of pretend-tools which would perform wonders if ever a set of facts should turn up in the right form" (G. D. N. Worswick, 1972, p. 79). Our job as teachers is to assist undergraduate students in gaining an understanding of operational tools and techniques while they learn to appreciate the limitations of the procedures and their legitimate use in statistical inference.

REFERENCES

Becker, William E. and Harnett, Donald L., *Business and Economics Statistics With Computer Applications*, Reading: Addison Wesley, 1987.

______ **and Waldman, Donald M.,** "The Probit Model," in William Becker and William Walstad, eds., *Econometric Modeling in Economic Education Research*, Boston: Kluwer-Nijhoff, 1987, 135–40.

Bryan, Marvin, "Business Forecasting," *PC Magazine*, August 1986, *5*, 211–34.

Drèze, Jacques H., "Nonspecialist Teaching of Econometrics: A Personal Comment and Personalistic Lament," *Econometric Reviews*, No. 2, 1983, *2*, 291–99.

Fels, Rendigs, "Developing Independent Problem-Solving Skills in Economics," *American Economic Review Proceedings*, May 1974, *64*, 403–07.

______ **and Uhler, Robert G.,** *Casebook of Economic Problems and Policies*, St. Paul: West Publishing, 1st ed., 1974 (5th ed. edited by Fels and Stephen Buckles, 1981).

Hansen, W. Lee, "What Knowledge is Most Worth Knowing For Economics Majors?," *American Economic Review Proceedings*, May 1986, *76*, 149–52.

Johnston, J., *Econometric Methods*, New York: McGraw Hill, 3rd ed., 1984.

Rose, E. L., Machak, J. A. and Spivey, W. A., "A Survey of the Teaching of Statistics in Business Schools," unpublished paper, Graduate School of Business Administration, University of Michigan, October 1986.

Seigfried, John J. and Wilkinson, James T., "The Economics Curriculum in the United States," *American Economic Review Proceedings*, May 1982, *72*, 125–38.

Sowey, Eric R., "University Teaching of Econometrics: A Personal View," *Econometric Reviews*, No. 2, 1983, *2*, 255–89.

Worswick, G. D. N., "Is Progress in Economic Science Possible?," *Economic Journal*, March 1972, *82*, 73–86.

Coping with the Diversity of Student Aptitudes and Interests

By VIJAYA G. DUGGAL*

There is no dispute in the economic profession that a background in quantitative methods is essential for undergraduates majoring in economics. What may not be so clear is what is expected of the student who has completed the requirements in quantitative methods. Is the student expected to be able to advance the discipline of statistics and econometrics, or is he or she merely expected to be able to use the tools to find quantitative solutions to issues he might face in the profession? When the question is posed in this manner, the majority will consent to the second expectation. Do we, in the quantitative courses we teach, prepare them adequately to fulfill this expectation, or are we as instructors secretly driven to expose our students to all mathematical derivations that we studied in graduate school?

If we dealt with only the brightest of students in an average university, it would not matter how we approached the subject, for these students have the ability to comprehend and absorb whatever they are taught. For the average student, however, it is important to limit ourselves to what we think should be their essential repertoire.

This paper is not intended to cover the teaching of quantitative methods in a comprehensive manner, or to provide any systematic revaluation. Instead, it focuses on some changes in our teaching methods that may enhance certain segments of a quantitative course. Essentially, I find that we spend too much time on statistical theory on the one hand, and hand calculations on the other, at the expense of conceptual understanding of the material. Our method of teaching needs to be modified more than it has been in order to exploit advances in computer technology. There is a need for greater emphasis on the relevance of the material and practical applications of the tools we present.

I. Making the Material Relevant in the Beginning

The undergraduate who likes the magic of mathematics or indulges in learning for learning's sake, gets engrossed in the material as it is presented, and does not need any additional motivation. The average student, however, working his (her) way through calculations for regression coefficients and standard errors does not see beyond the calculations, and remains disinterested. In order to get the student to even make an attempt to understand the subject matter, an important hurdle to cross is that of motivation. Motivation, I believe should be forthcoming with relevance. Students may see the relevance of quantitative methods if they can be made to appreciate early on in the course the practical applications of what they are about to study. Most students starting their course in quantitative methods have no idea of how it will be useful to them.

Studies have shown that there is increased learning and longer-term retention of what is learned when there is active participation on the part of the students. The earlier in the course such active participation takes place, the greater is the benefit.

An interesting way to motivate the students and simultaneously get them actively involved is to expose them in the beginning of the course to a finished product of econometric investigation such as a model and have them work with it to find answers to issues of current policy. After the students have had a chance to see the relevance, they are eager to work backwards to see how a model is put together and what problems are encountered at each stage of development.

This approach would have been unfeasible to use for all undergraduates majoring in economics in the past because of its heavy dependence on mainframe computers. Time

*Associate Professor of Economics, Widener University, Chester, PA 19013. I am deeply indebted to F. Gerard Adams and Lawrence R. Klein for constructive comments.

sharing on mainframes has always been expensive and computer demonstrations in the classroom were rarely arranged because of the inconvenience of slow response time. With the arrival of personal computers and the increase in their capability to handle fairly complex systems of equations and their inherent mobility, we have access to resources whose potential we have barely begun to tap.

To take an example of a finished product of econometric investigation, let us consider a macro model of the U.S. economy—a model that has both demand and supply fully integrated and which has a fair number of fiscal and monetary tools embedded in it. The possible candidates representing Keynesian doctrine are the Wharton Quarterly model, the Michigan model, and the Fair model. We could just as well take a model based on monetarism, or on the rational expectations theory. The purpose is not to make the students experts on any particular model, but rather to show them how a model relates to the subject matter of economics and how it can be useful in solving problems.

The model that I have used in this way is the Wharton Mini Growth Model which is a skeletal version of the Wharton Long Term Model stripped of its industry detail. The only industry detail kept was that of energy, which dates the model as having been developed in the mid to late 1970's. There are behavioral equations for consumption, business investment, residential investment, imports, exports, the price level, wage rate, employment, man-hours, and interest rates. Tax collections are endogenously estimated using exogenous tax rates and relevant incomes as independent variables. Government spending is exogenous in real terms. Actual spending varies with the price level. A large number of identities make it possible to look at variables within the national income accounting framework.

The model is solved for the last ten years of history by adjusting each equation of the model by the corresponding single-equation residuals. The path simulated by the model using constant adjustments in this manner tracks historical values accurately. This allows us to ignore the inherent problems of errors in forecasting and to start with a realistic solution that students can relate to. The base solution is stored and used as a standard of reference to which all other solutions generated in and out of the classroom are compared.

After an introductory presentation of a global view of modeling and its potential use in studying implications of possible policy changes, the students should see a classroom demonstration of the effect of a change in some exogenous variables such as defense expenditures. A new solution can be generated with the same exogenous assumptions and constant adjustments as the base solution except for a $10 billion increase in defense spending for every period. This requires retrieving the base solution and increasing defense spending by $10 billion. The new solution can be compared with the old over time. Its impact on GNP, immediate and long run, is sure to remind the students about government multipliers they studied in their macro course. They can quantitatively build up the multiplier from its components and trace the leakages in the system. The model which has been nothing but a black box for the students begins to take form in their minds as they investigate how the equations in the model represent the behavior of the economy.

Additional simulations with the model to study the sensitivity of the major economic variables to a decrease in personal income taxes and to a balanced increase in expenditures and taxes reinforces in the mind of the student the utility of modeling and therefore of statistical analysis.

The above can be completed in the first three to four hours of class time. For the students to really get a feeling for the usefulness of modeling, they must work with it independently. Assignments can be of varying degrees of complexity: 1) changing spending or taxes to achieve a target path of unemployment rate; 2) changing monetary policy to target inflation rates to a given level; and 3) changing both monetary and fiscal policy instruments so as to meet given targets on both inflation and unemployment rate. The students should report the results of their exercise and on their success and frustrations as they work on it. Some of the bright students may even begin questioning the specifications of the equations of the model.

The students by this time may or may not have mastered the equations of the model and the linkages within it. However, they certainly would have acquired a feeling for the relevance of econometric analysis. In addition, they would have received all the benefits of actively participating in the learning process at the onset of the course.

The end product of econometric investigation used in the manner indicated above does not have to be a macro model. I used it as an example because I feel comfortable dealing with it due to my particular background and also because I feel that the macro issues which it can address really excite the students. What is important is that some final product of econometric investigation gets discussed and played with by the students in the beginning of the course, rather than at its end where one would naturally think of fitting it.

This reversal of the natural sequence requires more than a proportionate effort on the part of the instructor. The students are being given a handful to learn and to do, and they are going to require more of the instructor's attention in the early part of the course. That in itself should have additional benefit. As Michael Wetzstein and Josef Broder pointed out, "The marginal effectiveness varies through time for a given level of [instructor] characteristics. The additional teaching effectiveness is greatest in the beginning part of the course,..." (1985, p. 57).

The above is an initial heavy dose of relevance. To sustain the momentum thus created, the problems given to them with each new concept must be economically significant. The nine laboratory experiments in econometrics described by Robert McNown and Gary Hunt (1984) which illustrate different statistical problems or econometric techniques would be very useful in this context.

II. The Use of Computer Demonstration in the Classroom

Most textbooks in their discussion of estimation techniques start with made up data on variables X and Y, because of the ease with which such data lends itself to hand calculations on the blackboard. It is far more preferable to take actual historical data the students have already been examining. There is plenty of software available which has the facility of storing large amounts of data and which has an estimation and simulation package. A demonstration in class of the scatter diagram of the actual values of consumption and disposable income, for example, and then getting the computer to generate a least squares line, leaves a strong visual image with the student. No amount of placing dots on the blackboard and chalking a line through them creates the impression either of the least squares principle or of the power of the computer as does one classroom demonstration. It is good to have this demonstration precede the theoretical or mathematical discussion of the least squares estimation procedure.

III. Conceptual Understanding

The fact that the least squares line minimizes the square of the vertical distance between the points and the line is, no doubt, of the utmost importance. If deriving the values of the regression coefficients mathematically results in a more lasting impression, reinforcing the fact of minimization, then, of course, these should be derived. But frequently I see students so bogged down in understanding the algebra and calculus that they have lost the significance of what they were doing, and have no intuitive understanding of the concepts. The bright student will absorb and retain both the concept and the mechanics, but for the less fortunate, the conceptual understanding should take priority. Theory and mathematics have to play second fiddle to such an understanding if econometrics is to be required of all undergraduates majoring in economics, and if we wish our students to apply the skills they learn appropriately when the situation demands. Some universities offer two econometrics courses, one more mathematical than the other, to accommodate differing aptitudes.

I contend that more than three-fourths of students finishing a course in quantitative methods who know the mechanics of hypothesis testing have no conceptual comprehension of the sampling distribution of the mean. Extracting all possible samples of size 2 from a population of 4 and showing the relationship between the population statistics and the sample statistics, and deriving the entire sampling distribution of the mean, although contrived, is an exercise well worth the time for the understanding it provides of the sampling distributions and the central limit theorem.

The emphasis on understanding of the meaning of the concepts and their significance dictates the choice of the formulas to be discussed in class. The formulas no longer have to be those that are expedient from the calculation point of view but the ones that go to the core of the meaning. For example,

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\Sigma(Y-\bar{Y})^2}}$$

although hopelessly cumbersome from a calculation point of view, is superior in the interpretative meaning of the correlation coefficient than any of the competing formulas. Two simple examples of perfect positive and negative correlation to be done by hand using this version of the formula drives home its meaning better than long laborious hand calculations with semirealistic data using alternative computationally expedient formulas.

In our attempts to make use of computers and redirect what we emphasize, we need to be careful not to deemphasize hand calculations altogether. Hand calculations certainly do not have to be done on real data, for that is the primary raison d'etre for computers, but hand calculations are absolutely essential using carefully constructed data with the version of the formula that makes the meaning of the concept evident.

IV. Independent Research

Towards the end of the course, students should undertake independent study that allows them to apply what they have learned. Two possibilities come to mind.

Each student could pick an equation from the model that was used in the beginning and improve upon it by changing specification—adding new variables or altering functional form. He or she should replace the original equation in the model by the one he estimates and test for the error and simulation properties of the modified model.

Students could be divided in groups of 4 or 5. Each member of the group contributes an equation that the group puts together as a simultaneous system explaining some economic phenomena. The group then tests for the error and simulation properties of the model it has constructed.

In either of these exercises, the students should be encouraged to collect at least some of their own data. In this way, the students are forced to know something about the nature and sources of the data and they begin to appreciate the strength and weakness of the data they use.

V. Conclusion

In order to maintain the interest level of the student enough to persuade him to put in the hard work necessary to master techniques of quantitative methods, it is essential to motivate him in the very beginning by showing him the relevance of what he is about to study. It is necessary to keep the relevance issue uppermost as different concepts are presented and discussed. We need to make more use of computer demonstration in the classroom because of the impact it has on the students. We need to choose formulas for concepts that explain the meaning of the term best, rather than those that are computationally expedient. Because the computer can do the laborious calculations, problems for hand calculations need to be devised solely from the point of view of clarifying the meaning of the concept.

We no longer need to subject our students to the drudgery of hand calculations of semirealistic problems. Instead, that time can now be used in studying the practical situations in which a particular concept can be productively used. Many more problems can be solved using the computer and a larger fraction of time can be left for the interpretation of results.

REFERENCES

McNown, Robert F. and Hunt, Gary L., "An Econometrics Laboratory," *The Journal of Economic Education*, Winter 1984, *15*, 71–76.

Sheinin, Yacov, "Wharton Mini Growth Model," Wharton Econometrics Working Paper, 1981.

Wetzstein, Michael E., and Broder, Josef M., "The Economics of Effective Teaching," *The Journal of Economic Education*, Winter 1985, *16*, 52–59.

FINANCE AND ECONOMICS[†]

The Interrelations of Finance and Economics: Theoretical Perspectives

By STEPHEN A. ROSS*

It is traditional in a discussion piece to organize the material in one of two ways. The writer can either take a historical perspective and attempt to explain how it is we got where we are today and where we are likely to go from here, or the writer can describe the current state of the art, dwelling on particular points of interest or promise in the prevailing research. Having quite recently done both, I thought I would take a somewhat different approach. I would like to try to briefly describe the main characteristics of a neoclassical theory of finance that captures the essential themes of modern finance and relate these characteristics to the general themes of economics.

Finance uses the modeling framework constructed in economics but, within this scaffolding, finance has taken a different methodological perspective. It is wrong to characterize finance, or financial economics to be formal, as simply another of the specialty areas of economics—not unlike, for example, labor economics or development economics or public finance. While finance is specialized in its focus on the financial markets, the differences between economics and finance only begin there. The principal distinction is one of methodology rather than of focus. If labor markets behaved like financial markets, the theories of finance would be used to study them. Indeed, the line where financial theoretic analysis leaves off and more conventional patterns of economic reasoning begin is an active research issue.

I. Data and Theory

In finance, the data are voluminous and of high quality. We have daily and even intradaily price data on the most important financial markets. Furthermore, the data are generated by processes that make it true transactions data or, at the least, close to that. The quality and the volume of the data subtly alter the reward structure for researchers in finance from that of other areas of economics. There is a premium on modeling close to the data. Which is not to say that there is no interest in the indicative models that are stylistic depictions of economic phenomena, rather than there is a great reward in explaining regularities in the existing data. That, in turn, leads to models whose variables are themselves observables rather than abstractions of classes of observables. There are very few models of securities markets with a variable called "all stocks."

Furthermore, there is a strong and subtle pressure to build models that utilize the data within the financial database. This aids and abets the focus on relative pricing. As financial economists we are concerned with the relation between the prices of different financial assets, rather than with their relation with other economic variables such as wage rates. This concern arises at least as much because of the fact that comparable data are available on different financial assets as because we believe that they trade in essentially the same market. Furthermore, the data are largely data on pricing; our volume data

[†] *Discussants*: William Brock, University of Wisconsin-Madison; Myron S. Scholes, Stanford University; Lawrence H. Summers, Harvard University.

*Yale School of Management, Yale University, New Haven, CT 06520. The thoughts expressed in this paper are my own, but if they have any value to them, it is the consequence of colleagues too numerous to mention.

are far rougher. It should be no surprise, then, that much of our theory is a theory of inelastic supply, and of price determination.

There is, of course, a chicken and egg issue at work here—an equilibrium problem, to be technical. Perhaps we collect the price data because those are the subject of our theories. I think not. I believe the reason we collect the price data is partly because they are available—like a high mountain, "it is there"—and partly for the intellectual reason that the financial markets are extremely liquid and as close to our purely competitive ideal as one can find in the real world. In such an environment, prices determine actions and quantities are secondary.

II. The Economic Approach and the Financial Approach

The apparatus of demand and supply and the attendant notions of equilibrium remain the major tools of economics. This is the framework that the economist uses to develop intuitions for situations as disparate as the holding of currency by the public and the workings of markets with two dominant suppliers. The methodology remains that of supply and demand, no matter how complex the information structure and no matter how intricate or arbitrary is the notion of equilibrium. Models are nearly always closed by setting price (or prices) so that supply matches demand.

More important than the particular model, though, is the intuition that underlies and motivates it. Whatever the market, the demand curve is positioned by external forces, such as preferences, and the supply curve is set by technology, and if the price is not at the equilibrium, familiar if not entirely specified forces are called into play to rectify matters. The supplier responds to a price above the equilibrium price by producing more than consumers will absorb, and some unseen friend of Walras responds to this excess of supply over demand by lowering the prices. Even game-theoretic models, such as those describing duopoly, can be viewed as extending this apparatus by specifying reaction curves of quantity-price responses that lead to the supply and demand equilibrium.

Paul Samuelson's textbook on economics has the following anonymous quote, "You can make even a parrot into a learned political economist—all he must learn are the two words 'supply' and 'demand'."

By contrast, the intuition of neoclassical finance is quite different. The focus of finance is micro theoretic and the intuition of finance is the absence of arbitrage. To make the parrot into a learned financial economist, he only needs to learn the single word "arbitrage."

This is not to say that the intuition and the theories of finance cannot be fit into the framework of supply and demand, rather that doing so does not gain us much. The fit is awkward and irrelevant at best. The ordinary demand and supply curves in competitive economies are drawn under the traditional assumption that other prices are held constant. In neoclassical finance the resulting demand curves are horizontal and perfectly elastic and the supply curves are either perfectly elastic or perfectly inelastic, depending on the problem being studied. What matters in such a situation is not movements along the curves in response to changes in price—such "responses" are unbounded—but, rather, where the curves are in the price-quantity picture. Unlike what occurs when elasticities are in the normal ranges, *everything* of interest is underneath the supply and demand picture and the picture is meaningless.

The forces of supply and demand have no meaning, since if the price is not the equilibrium price, then the difference between supply and demand is infinite. This is precisely what is meant by an arbitrage situation, and it is so qualitatively different from the economist's usual picture of demand and supply as to require a different approach.

The demand curves are perfectly elastic because of the implicit assumption that financial markets are filled with assets which are very close substitutes for one another. In the stock market, any one stock is characterized by its sensitivities or *betas* on innovations in the state variables that systematically affect returns. Diversification removes any contribution to an optimal portfolio's returns that comes from idiosyncratic forces which affect an individual stock's returns and that leaves only a stock's *betas* to influence the uncertain portion of the portfolio return. It follows that a stock is perfectly substituted by any other stock or any portfolio with the same pattern of *betas*.

In the option markets, the existence of close substitutes is the centerpiece of the entire pricing theory. Under some hypothesized circumstances, a derivative asset, that is, an asset whose return derives from that of another more primitive asset that underlies it, will be perfectly substituted for by a portfolio of the primitive asset and another asset such as a bond. It is worth looking more closely at this familiar situation to see how it can be fitted into the demand and supply framework. (The interested reader can pursue the following approach to option pricing more closely in John Cox, Ross, and Mark Rubinstein, 1979.)

Suppose that a stock pays no dividend in the period we are looking at and that its price follows the simple binomial model,

$$S_t\begin{cases} S_{t+1}=aS \\ S_{t+1}=bS \end{cases},$$

where subscripts denote time and where we will let $a > b > 0$, and refer to the move to aS as an up move and the move to bS as a down move.

A call option written on this stock is a derivative security whose payoffs are determined by the value of the stock. For example, if the call option matures next period, then its value next period will be given by $\max(S_{t+1}-K,0)$, where K is the exercise price of the call. If we assume that the call is "in the money" on an up move and "out of the money" on a down move, then the call will go in value from C_t to

$$C_t\begin{cases} C_{t+1}=aS-K \\ C_{t+1}=0 \end{cases}$$

Lastly we will introduce a riskless bond into this world that pays off at an interest rate of r no matter whether the stock goes up or down.

We now have enough information to construct the excess demand curve for the call option as a function of its price, C_t. As is usual in such a construction we will take other prices as given, notably S_t and the interest rate, r. (To prevent either the stock or the bond from dominating the other, we must have $a > 1+r > b$.)

Consider a one dollar investment in a portfolio of the bond and the stock that has α dollars invested in the stock and $(1-\alpha)$ invested in the riskless asset. The value of the dollar next period will be given by

$$1\begin{cases} \alpha a+(1-\alpha)(1+r) \\ \alpha b+(1-\alpha)(1+r) \end{cases}$$

in the two possible states of nature, a and b. (Notice that an investment of α in the stock at time t purchases α/S_t shares of stock, and that at time $t+1$ the investment will be worth $(\alpha/S_t)S_{t+1}=\alpha(S_{t+1}/S_t)$, or αa if the stock goes up and αb if the stock goes down).

Now, pick the investment in the risky asset, α, to be such that the return on the portfolio is the same as that on the call option if the stock goes down. In other words, set α such that

$$\alpha b+(1-\alpha)(1+r)=0,$$

or

$$\alpha=(1+r)/((1+r)-b).$$

The return on the call option and the portfolio are now identical if the stock goes down to bS. If the stock rises to aS, the call will have a return per dollar invested in it of $(aS-K)/C_t$ compared with the portfolio's return of

$$\alpha a+(1-\alpha)(1-r)=\left[\frac{(1+r)}{(1+r)-b}\right]a + \left[\frac{-b}{(1+r)-b}\right](1+r).$$

It follows that the return on the portfolio will exceed or fall short of that on the call option whenever the price of the call,

$$C_t \gtrless C_t^*,$$

where

$$C_t^* \equiv (aS-K)\left\{\left[\frac{(1+r)}{(1+r)-b}\right]a + \left[\frac{-b}{(1+r)-b}\right](1+r)\right\}^{-1}$$

If the stock goes down, then a one dollar investment in the call option will result in the same return, namely, a loss of the total investment, as a one dollar investment in the portfolio. If C_t is greater than C_t^*, and if the stock goes up, then the return on the call option will be less than the return on the portfolio. In such a circumstance, no investor will want to own the call option and the demand for it will be zero. In fact, the situation is more extreme than that.

If C_t is greater than C_t^*, then not only will investors not wish to hold the call, they will want to write calls, that is, sell them. Furthermore, since they can lock in a riskless arbitrage profit from doing so, there is no natural limit to their supply of calls. Investors can merely write a call, receive the value, C_t, and invest C_t^* dollars in the portfolio to hedge their sale. The call and the portfolio position will both be worthless in the down state, and the portfolio will produce $aS-K$ dollars in the up state which will exactly offset the investor's obligation to the call purchaser. With no liability, then, and with no investment, the investor has made an instant gain of $C_t - C_t^*$.

Similarly, if C_t is less than C_t^*, then the portfolio is dominated by the call option and investors can realize an arbitrage gain by shorting the portfolio and purchasing the call. The result will be an infinite demand for the call option. In other words, the excess demand curve for the call option is perfectly elastic at a price of C_t^*. There is an infinite demand for the call if its price is less than C_t^* and an infinite supply of the call if its price is above C_t^*.

This situation occurs because of the ability to construct a portfolio that is a perfect substitute for the call option.

III. Intuition and Theory

The intuition and the theory of finance are coconspirators. The theory is less a formal mathematical structure driven by its own imperatives—although that is always a danger—as it is a handmaiden that attempts to bridge the gap between the intuitions and the data. Nowhere is this role clearer than in the area of efficient markets.

Intuition tells us that an efficient market is one where all of the publicly and cheaply available information is used by investors to determine the values at which securities trade in the market. This means that the prices should "reflect" this information in some sense. It also means that an investor who simply makes use of this information should not be able to earn "abnormal" profits by doing so. In other words, trading schemes are doomed. Such is the basic intuition of efficient markets.

Turning this intuition into formal theory, though, and bringing it to the data is another matter. Actually, the problem of explaining the data is too important to wait for theory to establish the rigorous hypotheses for empirical analysis. Instead, the whole process becomes sloppier, and the intuitions themselves are used as interpretive guides in simple and straightforward empirical tests. Thus, the researcher tests whether rates of return are serially correlated without ever formally examining if that is a consequence of efficient markets generally and the inability to earn abnormal profits specifically.

There is nothing wrong with all of this. On the contrary, without it our intuition about the role of information in financial markets would still be unhoned and our theories would probably be even more rudderless than they are now. But, it does lead to a confusion between theory and testing that may be worth addressing.

Efficient market theories are perhaps the central area of this confusion. The intuition underlying the efficient market theories is the intuition of the lack of arbitrage. Just as an arbitrage opportunity occurs at a moment of time when, say, two different riskless interest rates prevail, intuition suggests that an arbitrage also can occur at two separated moments of time. It is this notion that drives the formulation of efficient market hypotheses.

IV. Information Economics and Finance

When we turn to the task of making this intuition explicit, the apparatus for doing so becomes the modern economics of information. In doing so, though, we run the risk of losing the arbitrage intuition of neoclassical finance. For example, it has become a truism in finance that empirical tests of efficient market hypotheses are joint tests of asset pricing models and efficiency. This jointness has become so accepted that it has taken efficient market tests out of the realm of arbitrage and, in large measure, made them indistinguishable from more traditional tests of asset pricing models.

I am less convinced of the truth of this truism than I used to be, and I see the same dilution of the arbitrage intuition in many applications of the economic theory of information to problems of modeling financial markets.

Bidding models have become the standard approach to developing a formal theory of mergers and acquisitions. Signalling models are now a familiar approach to the determination of financial structure. Agency models are formalized to develop theories of the separation of ownership and control in firms. Equilibrium models with information conveyed by sufficient statistic prices are the tool we use to understand trading in markets.

All of these approaches have enriched our understanding of a variety of phenomena in the financial markets, from the pricing of new equity issues to the rise of the "White Knight." But, I have the feeling that some backtracking has to be done to recover the intuition that began the whole process in the first place.

After all, finance has progressed very far by having a faith—some would say religious but I prefer to think of it as a proven first-order approach to problems—in the broad efficiency of markets. By and large, it is very difficult to "beat" the market (whatever that means) and somehow the current generation of information theory models too easily stray far from the original intuition of efficiency. This may be the inevitable consequence of looking at details more closely and at greater theoretical magnification, but I suspect that it is more the consequence of straying further from the data.

V. Corporate Finance

Nowhere have we strayed as far as in the area of the theory to the firm. Many of our theories are now indistinguishable from those of the transactional approach to the theory of the firm. Agency theory, be it informal and in the verbal tradition or the formal neoclassical models of the agency and moral hazard literature, is now the central approach to the theory of managerial behavior. Set aside is the original intuition of neoclassical finance that an arbitrage exists whenever a firm is mismanaged. This is not to say that this theme is missing from the present literature, but, rather, to express a personal view that it is receiving short shrift.

Perhaps this is appropriate and will lead to a better understanding of these matters. But, I get uncomfortable with large-scale game theoretic models of firm behavior in incomplete markets that are unmotivated and divorced from the financial setting that they purport to study. Such models have yet to produce a significant new idea or intuition in finance and insofar as they might just as well be models of the milk market as the financial markets, our expectations should not be very high.

VI. Conclusion

As I read over this piece I find that it sounds harsher than I feel. But, at the risk of continuing in the same vein I'll end on a heretical note. There is a great deal of discussion nowadays about bridging the gap between economics and finance. To some extent this is motivated by the well-intentioned and obvious view that each has something to offer the other.

But, contrary to this trend, I believe that it would be productive to maintain some distance between the two areas. Clearly, financial theorists should master modern economic theory and look to apply it to problems of interest in finance. Similarly, economics, in general, will greatly benefit from the tools and data developed in finance. An argument can be made that the intuition and early work on efficient markets was the impetus if not the cornerstone of the new neoclassical, rational expectations school of macroeconomics. Surely, too, the new financial tools for looking at financial market data will greatly enrich our understanding of how economies work.

But, much of what finance has accomplished and contributed to economics has been the result of working in a somewhat isolated and eccentric tradition. To the extent to which finance is successfully integrated into economics, this competing and successful strain may be bred out. Of course, without the continuing need to communicate with and satisfy the standards of mainstream economics, another danger arises. By the standards of mainstream modern medicine, chiropractors are also eccentrics and a bit more integration might have done them some good, not to mention their patients.

This risk seems to me worth running and the past record of the friendly competition between economics and finance has been extraordinary. To mention just two of the results, finance gave economics its penchant for rational expectation, and it has now given it option pricing and the general arbitrage theory.

REFERENCES

Cox, John C., Ross, Stephen A. and Rubinstein, Mark, "Option Pricing: A Simplified Approach," *Journal of Financial Economics*, September 1979, 7, 229–63.

Samuelson, Paul, *Economics*, *Ninth Edition*, New York McGraw-Hill, 1973, p. 58.

The Interrelations of Finance and Economics: Empirical Perspectives

By MICHAEL R. GIBBONS*

The title of this paper is somewhat inappropriate, for it may suggest that finance is a study separate from economics. In fact, most researchers in finance refer to themselves as financial *economists*, and many have done their graduate work in departments of economics. A more appropriate (but longer) title would refer to the interrelations of financial economics and other fields in economics. Given finance is a field within economics, it is not surprising that finance has borrowed heavily from other disciplines within economics and that the reverse has occurred as well—although the latter is a newer phenomenon than the former.

Financial economics has a long tradition of empirical work which will be the focus of this paper. I categorize the cross fertilization between finance and economics in the next four sections. The first section discusses the sharing of econometric methods. Section II focuses on situations where other fields in economics also attempt to explain prices of financial securities. Because of the quality and quantity of financial data, finance has served as an empirical laboratory for other fields in economics; this is discussed in the third section. Finally, since security prices are governed in part by expectations about future economic variables, there have been attempts to extract these expectations (as well as other unobservables) from the observed prices of financial assets. Section IV notes some examples of where unobservables have been extracted from prices of securities.

Using financial data to measure the economic impact of certain events or to extract unobservables presumes that the participants in the financial market are rational. To the extent that this rationality assumption is violated calls into question the usefulness of financial data. Section V discusses some of the recent empirical anomalies in financial economics as well as their ramifications for employing financial data.

This paper is not an exhaustive survey of all the interrelations between finance and economics. Instead the paper only attempts to illustrate some of the interrelations by relying on a few examples.

I. Interrelations Due to Sharing of Econometric Methods

There are several examples where methodology has been shared between finance and other fields of economics. The methodologies for "event studies" and for tests of asset pricing models are two important illustrations.

The event study methodology, pioneered by Eugene Fama et al. (1969), was developed to measure the impact of certain events (for example, the announcement of stock splits) on security prices. The method is useful in describing the reaction of security prices as well as testing various hypotheses relating to certain public announcements. Variants of this original methodology has been used extensively in the finance literature, and usually the stated purpose of these articles has been to test the "efficient markets hypothesis" as defined by Fama (1970). As discussed in Section III below, this method along with financial data has been employed in various fields of economics, including the economics of information, economics of regulation, industrial organization, and macroeconomics.

The event study methodology provides an example where a technique was developed in financial economics and then migrated to other areas in economics. In contrast, more

*Graduate School of Business, Stanford University, Stanford, CA 94305. In preparing this paper I had useful conversations with Ed Lazear, Paul Pfleiderer, Peter Reiss, and Myron Scholes. I am grateful for their help, but of course I remain responsible for any errors and oversights. Financial support was provided in part by the Stanford Program in Finance.

recent tests of asset pricing models are largely an outgrowth of econometric procedures that were developed in macroeconomics for testing cross-equation restrictions implied by models of rational expectations. My 1982 paper developed a framework for examining the cross-equation restrictions implicit in asset pricing models. In some ways, the econometrics for studying the cross-equation restrictions implicit in financial models are easier than those in macro models. Typically the restrictions from financial models are already stated in terms of the parameters of the reduced form, the underlying dynamics are simple in that serial correlation can usually be ignored, and normality may be a reasonable distributional assumption. These simplifications have allowed financial economists to provide an analysis of power to reject and more intuition about these tests of cross-equation restrictions (for example, my paper with Stephen Ross and J. Shanken, 1986) than we usually find in the macro literature.

There is little doubt that the sharing of econometric methods between finance and other areas in economics will continue. Two areas of potential growth are in the application of Lars Peter Hansen's (1982) generalized method of moments (GMM) and the analysis of temporal aggregation bias in econometric models. GMM has already been used to estimate asset pricing models (for example, Hansen and Kenneth Singleton, 1982, and D. Brown and myself, 1985), and I expect even more applications in the future because the approach is well suited to models of interest to financial economists.[1] These future applications should deepen our understanding of GMM as an econometric technique.

Temporal aggregation bias is a second area of future sharing of econometric methods. Many financial models are derived in a continuous time setting because of theoretical tractability. Yet, observations are available only over discrete sampling intervals. Thus, financial models provide econometricians with a perfect setting for studying temporal aggregation bias.

II. Interrelations Due to a Common Interest in Explaining Asset Prices

Macroeconomics has a long standing interest in explaining prices of financial assets, especially government bonds; yet, financial models and macroeconomic models have surprisingly little in common. Part of this discrepancy arises from a difference in focus between the two fields. Quite naturally, macroeconomists are interested in how policy variables (for example, money supply) affect interest rates while financial economists focus on perfectly competitive and complete markets where real economic variables (for example, consumption) play a critical role. (For example, see the empirical work by D. Breeden et al., 1986; K. Dunn and Singleton, 1986; and Hansen and Singleton.)

Not only do the explanatory variables differ between finance and macro, but also the dependent variables are not exactly the same. Traditional macro models refer to *the* rate of interest and typically ignore other assets like equity. In contrast, financial models not only have many types of assets but also various rates of interest depending on the maturity of the bond. While financial models tend to be richer in terms of their implications across assets, macro models tend to give more attention to the implications across time for a given asset. These characterizations of the literatures on finance and macroeconomics certainly do not apply to all the empirical work in either field. No doubt counterexamples could be provided. Yet, I think these crude characterization do convey a sense of each literature.

Given both fields have a common interest in asset prices but a different focus, one would think each field could profit by understanding the other. I believe that macro models have incorporated more ideas from finance than financial models have from macro. One interpretation of this behavior is that finance has more to offer than macro; however, I believe finance still has a great

[1] For example, GMM can easily handle problems that arise from overlapping data which is a common problem in finance and economics (see Fama and myself, 1984, and Hansen and R. Hodrick, 1980).

deal to learn from macro. Let me begin by giving an example of how finance has affected macroeconomics.

The notion of a large and efficient capital market has affected how some monetary economists view the transmission mechanism between money supply and interest rates. Some of the early stories of how open market policy affects interest rates are inconsistent with the view that significant price pressure is not observed at the time of large block sale of financial securities documented in the finance literature by Myron Scholes (1972). After formulating an econometric model based on efficient markets, Frederic Mishkin (1981, 1982) rejects this "liquidity effect" from monetary policy; he finds little evidence to support the view that increases in monetary growth lead to declines in either short- or long-term interest rates. Similarly, Michael Rozeff (1974) finds that lagged values of money supply are not correlated with stock returns.

In contrast to macro, financial models rarely attempt to explain the time-series properties of asset prices in terms of other variables. As Lawrence Summers points out, mainstream finance has not "attempted to account for the stock market boom of the 1960s or the spectacular decline in real stock prices during the mid-1970s" (1985, p. 634). Similarly, financial economists have not tried to explain the dramatic changes in the real rates of interest that have been observed in the 1980's. The study of the properties of these time-series is a missed opportunity. While recent empirical work is attempting to rectify this oversight by modeling the changing conditional moments of the distribution of asset returns, it is difficult to understand why the work did not start sooner. I have only two explanations. First, financial theory has only recently provided satisfactory general equilibrium models where the stochastic process for asset prices is an endogenous outcome from assumptions about consumption and production (for example, see John Cox et al., 1985). Until recently, econometric analysis lacked theoretical guidance on how to model changing conditional moments in a multiperiod model of uncertainty. Second, I believe, but cannot prove, that a large number of empirical researchers in finance operated as if efficient markets require conditional expected returns to remain constant. Certainly, they paid lip service to the notion of a joint hypothesis of efficient markets and a correct equilibrium model, but most tests in practice assumed constant conditional expected returns as the equilibrium model. The uncovering of "anomalies" has forced empirical researchers to entertain the notion that conditional expected returns may change in an efficient market. Hopefully, future empirical research in finance will give as much attention to the equilibrium time-series properties of asset prices as past research has given to the cross-sectional characteristics.

Finance and macroeconomics are not the only fields which are interested in explaining asset prices. Public finance attempts to document the effect of taxes on security prices. James Poterba and Summers (1985) provide a nice discussion of both the finance and public finance literatures in terms of the effect of dividend taxation.

III. Interrelations Due to the Financial Database

Financial economists are fortunate because we have easy access to a large quantity of high-quality data. To the extent that financial data are relevant to other areas of economics, the data provide a good empirical laboratory to investigate hypotheses that are not necessarily an outgrowth from a study of financial theory.

Perhaps the best known example of this type of interconnection involves the economics of regulation. In this case not only have financial data been employed to study the impact of a change in regulation, but the event study methodology has been adopted as well. In an influential paper, G. William Schwert (1981) develops this methodology in the context of analyzing regulations. Examples of such applications of financial data and the event study methodology include a study on wage and price controls (R. Ruback, 1982), antitrust regulation (K. Schipper and R. Thompson, 1983), airline regulation (P. Spiller, 1983), product recalls (G. Jarrell and Sam Peltzman, 1985), motor carrier regulation (N. Rose, 1985), and government ownership of shares of publicly traded companies (C. Echel and T. Vermaelen, 1986). In many ways it is surprising that this type of study can detect statistically significant effects in stock returns from regulatory changes. While the data are of high quality, stock returns generally have high variance. Furthermore, these studies are typically plagued by small samples with an imprecise date for when the regulation may be incorporated into stock prices. Simulation evidence in S. Brown and J. Warner (1980) suggests that the inability to date changes in expectations in the security market lead to tests with little power, and J. Binder (1985) confirms this conclusion using a sample of events. Despite these problems, empirical research on the economics of regulation continues to document reactions by the stock market to changes in regulation; yet, I remain pessimistic about future empirical research in this area.

Financial economics has also developed a strong connection with the economics of information. Some of the theoretical results in the economics of information are difficult to test, especially if financial data were not available. Scholes provides a demonstration how the market reacts to informed traders in the context of large block sales. M. Wolfson (1985) shows that incentive problems and reputation effects can be examined empirically by studying oil and gas tax shelter programs. Thus, financial data offer a useful empirical laboratory.

Given financial data are easily available, it is not unusual for stylized empirical facts to get established first, and theoretical results follow to rationalize these facts. Since the theories are designed to fit the data, it is difficult to design clean tests of the models. Nevertheless, this style of research has led to further connection between finance and the economics of information. K. Rock (1986) uses adverse selection and winner's curse to provide a theoretical explanation for why initial public offerings are underpriced. Using a signalling equilibrium model, Milton Harris and Artur Raviv (1985) are able to justify existing empirical evidence concerning the call policy of convertible debt; their model justifies why calls are delayed and why we see negative stock returns at the call announcement. A. Shleifer and R. Vishny (1986) develop a model which rationalizes why stock prices of target firms decrease after the payment of greenmail. Of course, there is a wealth of existing empirical evidence about acquisitions and mergers. (Michael Jensen, 1983, provides a convenient summary of much of the evidence.) I expect future theoretical research will attempt to explain more of these stylized facts using game theory and bargaining models.

IV. Interrelations Due to Measurement of Unobservables

Since financial theory considers intertemporal behavior under uncertainty, the field naturally relates asset prices today to the characteristics of unknown future events. For example, asset prices are clearly affected by expectations about future inflation, interest rates, and earnings. These expectations are not directly observable, but in some cases they may be inverted from observable asset prices. Gathering information about these expectations is of great interest to economists in many areas, especially given the high quality of the financial data.

Fama (1975) and Fama-Gibbons extract measures of expected inflation from nominal interest rates using the Fisher equation. They find that these implicit expectations provide good forecasts of future inflation. In fact, the implicit expectations seem to dominate survey measures of expected inflation which are sometimes used in macroeconomic models.

Traditional hypotheses about the term structure of interest rates have fascinated both financial and nonfinancial economists alike. Usually these hypotheses are examined with respect to the predictive content of implicit forward rates. (For examples of this kind of empirical work, see Fama, 1976, 1984.) With the development of more modern theories of the term structure (for example, Cox et al., 1985), a deeper theoretical understanding of the biases in forward rates is now available and should lead to further work on the term structure by finance and other areas in economics. One area where finance has been particularly useful to empirical work in economics is in making the distinction between forward versus futures contracts (see Cox et al., 1981), which are sometimes used to forecast future interest rates rather than implicit forward contracts.

This section has focused on extracting measures of expectations from prices of assets. However, it is also possible to invert for other kinds of unobservables. For example, S. Brown and P. Dybvig (1986) suggest a way to estimate the model of the term structure due to Cox et al. (1985). This estimation scheme creates a time-series on an unobserved state variable (in this particular case, the instantaneous rate of interest) which could be used in other empirical work. The ability to use asset prices to get out unobservables will continue to be an important link between finance and economics. There are many aggregate series that economists would like to measure more precisely, and financial models offer some hope.

V. Recent Empirical Anomalies in Finance

There are a growing number of empirical anomalies in financial economics which cast some doubt on the usefulness of the exercises discussed in Sections III and IV above. If asset prices appear to be inconsistent with rational behavior, then they may not provide a good empirical laboratory for other fields in economics and a good set of measures for unobservable variables.

Everyone has a favorite set of anomalies. My favorite list includes the January effect (D. Keim, 1983), the Monday effect (Ken French, 1980, myself with P. Hess, 1981), discounts on closed-end mutual funds, and the volatility of stocks during trading vs. nontrading time periods (French and Richard Roll, 1986). To make matters worse, financial economist may be too optimistic about market efficiency based on past work using event studies. As Summers (1986) argues, these tests have little power to detect inefficiency due to the variability of stock returns, so failure to reject provides little comfort. Nevertheless, while there are many things that we do not understand about security markets, I think it is premature to conclude that asset prices are not useful for the exercises discussed above. For example, economists seem to be fascinated by the apparent excess volatility in stock returns as documented by Robert Shiller (1981); such evidence could have been used to reject rationality in security markets and the usefulness of financial data. However, A. Kleidon (1986) has made a convincing case that Shiller's conclusions are an artifact of inappropriate econometric methods.

I hope there will be increased interactions between finance and economics, for I think the potential payoffs are large.

REFERENCES

Binder, J., "Measuring the Effects of Regulation with Stock Price Data," *Rand Journal of Economics*, Summer 1985, *16*, 167–83.

Breeden, D., Gibbons, M., and Litzenberger, R., "Empirical Tests of the Consumption-Oriented CAPM," Research Paper No. 879, Graduate School of Business, Stanford University, 1986.

Brown, D. and Gibbons, M., "A Simple Econometric Approach for Utility-Based Asset Pricing Models," *Journal of Finance*, June 1985, *40*, 359–81.

Brown, S. and Dybvig, P., "The Empirical Implications of the Cox, Ingersoll, Ross Theory of the Term Structure of Interest Rates," *Journal of Finance*, July 1986, *41*, 616–28.

______ **and Warner, J.,** "Measuring Security Price Performance," *Journal of Financial Economics*, September 1980, *8*, 205–58.

Cox, J., Ingersoll, J., and Ross, S., "The Relation Between Forward Prices and Futures Prices," *Journal of Financial Economics*, December 1981, *9*, 321–46.

______, ______, **and** ______, "A Theory of the Term Structure of Interest Rates," *Econometrica*, March 1985, *53*, 385–407.

Dunn, K. and Singleton, K., "Modeling the Term Structure of Interest Rates Under Non-Separable Utility and Durability of Goods," *Journal of Financial Economics*, September 1986, *17*, 27–56.

Eckel, C. and Vermaelen, T., "Internal Regulation: The Effect of Government Ownership on the Value of the Firm," *Journal of Law and Economics*, October 1986, *29*, 381–404.

Fama, E., "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, May 1970, *25*, 383–417.

______, "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

______, "Forward Rates as Predictors of Future Spot Rates," *Journal of Financial Economics*, October 1976, *3*, 361–77.

______, "The Information in the Term Structure," *Journal of Financial Economics*, December 1984, *13*, 509–28.

______ **et al.** "The Adjustment of Stock Prices to New Information," *International Economic Review*, February 1969, *10*, 1–21.

______ **and Gibbons, M.,** "A Comparison of Inflation Forecasts," *Journal of Monetary Economics*, May 1984, *13*, 327–48.

French, K., "Stock Returns and the Weekend Effect," *Journal of Financial Economics*, March 1980, *8*, 55–70.

______ **and Roll, R.,** "Stock Return Variances: The Arrival of Information and the Reaction of Traders," *Journal of Financial Economics*, September 1986, *17*, 5–26.

Gibbons, M., "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, March 1982, *10*, 3–27.

______ **and Hess, P.,** "Day of the Week Effects and Asset Returns," *Journal of Business*, October 1981, *54*, 579–96.

______, **Ross, S., and Shanken, J.,** "A Test of the Efficiency of a Given Portfolio," Research paper No. 853, Graduate School of Business, Stanford University, 1986.

Hansen, L. P., "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, July 1982, *50*, 1029–84.

______ **and Hodrick, R.,** "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, October 1980, *88*, 829–53.

______ **and Singleton, K.,** "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, September 1982, *50*, 1269–86.

Harris, M. and Raviv, A., "A Sequential Signalling Model of Convertible Debt Call Policy," *Journal of Finance*, December 1985, *40*, 1263–82.

Jarrell, G. and Peltzman, S., "The Impact of Product Recalls on the Wealth of Sellers," *Journal of Political Economy*, June 1985, *93*, 512–36.

Jensen, M., "Symposium on the Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics*, April 1983, *11*, 3–472.

Keim, D., "Size-Related Anomalies and Stock Market Seasonality: Further Empirical Evidence," *Journal of Financial Economics*, June 1983, *12*, 13–32.

Kleidon, A., "Variance Bounds Tests and Stock Price Valuation Models," *Journal of Political Economy*, October 1986, *94*, 953–1001.

Mishkin, F., "Monetary Policy and Long-Term Interest Rates: An Efficient Markets Approach," *Journal of Monetary Economics*, January 1981, *7*, 29–55.

______, "Monetary Policy and Short-Term Interest Rates: An Efficient Markets-Rational Expectations Approach," *Journal of Finance*, March 1982, *37*, 63–72.

Poterba, J. and Summers, L., "The Economic Effects of Dividend Taxation," in E. Altman and M. Subrahmanyam, eds., *Recent Advances in Corporate Finance*, Homewood: Richard Irwin, 1985, ch. 9.

Rose, N., "The Incidence of Regulatory Rents in the Motor Carrier Industry," *Rand Journal of Economics*, Autumn 1985, *16*, 299–318.

Rock, K., "Why New Issues are Underpriced," *Journal of Financial Economics*, January/February 1986, *15*, 187–212.

Rozeff, M., "Money and Stock Prices: Market Efficiency and the Lag in the Effect of Monetary Policy," *Journal of Financial Economics*, September 1974, *1*, 245–302.

Ruback, R., "The Effect of Discretionary Price Control Decisions on Equity Values," *Journal of Financial Economics*, March 1982, *10*, 83–105.

Schipper, K. and Thompson, R., "The Impact of

Merger-Related Regulations on the Shareholders of Acquiring Firms," *Journal of Accounting Research*, Spring 1983, *21*, 184–221.

Scholes, M., "The Market for Securities: Substitution Versus Price Pressure and the Effects of Information on Share Price," *Journal of Business*, April 1972, *45*, 179–211.

Schwert, G. W., "Using Financial Data to Measure Effects of Regulation," *Journal of Law and Economics*, April 1981, *24*, 121–58.

Shiller, R., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *17*, 421–36.

Shleifer, A. and Vishny, R., "Greenmail, White Knights, and Shareholders' Interest," *Rand Journal of Economics*, Autumm 1986, *17*, 293–309.

Spiller, P., "The Differential Impact of Airline Regulation on Individual Firms and Markets: An Empirical Analysis," *Journal of Law and Economics*, October 1983, *26*, 655–90.

Summers, L., "On Economics and Finance," *Journal of Finance*, July 1985, *40*, 633–35.

______, "Does the Stock Market Rationally Reflect Fundamental Values," *Journal of Finance*, July 1986, *41*, 591–600.

Wolfson, M., "Empirical Evidence of Incentive Problems and their Mitigation in Oil and Gas Tax Shelter Program," in J. Pratt and R. Zeckhauser, eds., *Principals and Agents: The Structure of Business*, Boston: Harvard Business School Press, 1985, ch. 5.

DYNAMIC MODELS OF TAXATION[†]

A Dynamic Theory of Factor Taxation

By KENNETH L. JUDD*

Many important questions in macroeconomics concern the impact of taxation and spending policies on private resource allocation. One approach, typified by Robert Hall (1971) and William Brock and Stephen Turnovsky (1981), is to use dynamic general equilibrium models to address basic issues in fiscal and tax policy. The objective is to explicitly examine macroeconomic issues without embracing the market imperfections (fixed prices, missing markets, money illusion, etc.) found in conventional macroeconomic analysis.

There is currently great interest in dynamic fiscal policy problems. This decade has already seen two major adjustments in tax policy, and many attempts to substantially alter spending patterns. The result of this tumult has been unprecedented peacetime deficits and much uncertainty about what measures will ultimately be used to bring the budget into balance.

In this paper, I discuss the impact of alternative fiscal policies in a dynamic general equilibrium model. I examine the short-run effects of fiscal policy changes, the efficiency cost of alternative dynamic tax policies, the effects of uncertain policy formation, and the redistributive effects of factor taxation.

I. Model

I briefly sketch the details of the basic model, referring the reader to my paper (1987) for details. I assume an economy with a larger number of identical, infinitely lived individuals with a dynamic utility function

$$U=\int_0^{\infty} e^{-\rho t}u(c,l)\,dt+\int_0^{\infty} e^{-\rho t}v(g)\,dt,$$

where ρ is the pure rate of time preference, $c(t)$ is the rate of consumption of private goods, $g(t)$ is public consumption, and $l(t)$ is labor supply at t, and $u(c,l)$ and $v(g)$ are felicity flows from private and public consumption, respectively. I assume one capital stock, which depreciates at a constant rate of δ. The K units of capital together with labor supply l, both per capita, produces gross output at the rate of $F(K,l)$ per capita.

I assume a simple tax structure where the marginal tax rate is τ_K for capital income net of depreciation and τ_L for labor income, and where firms receive an investment tax credit (ITC) for gross investment. Individuals may receive a lump sum subsidy, representing both nonproportionality in the income tax and public provision of goods which are perfect substitutes for private goods. Government also issues bonds, allowing taxes at one time to finance government consumption at another. Such bonds are perfect substitutes for private capital.

First, I must describe individual choices in such a dynamic world. They face a before-tax, but net of depreciation, return on physical assets of $r(t)$ and a wage of $w(t)$. If λ represents the private marginal value of capital, optimality implies that λ obeys the Euler equation,

$$\dot{\lambda}=\lambda\left(\rho-\left(r(1-\tau_K)+\delta\theta\right)/(1-\theta)\right),$$

and that consumption demand and labor supply satisfy

$$u_c(c,l)=\lambda=u_l(c,l)/\left(w(1-\tau_L)\right).$$

In equilibrium, $r=F_K$ and $w=F_l$. Combining these conditions with the Euler equation,

[†] *Discussants*: James M. Poterba, Massachusetts Institute of Technology; B. Douglas Berhnheim, Stanford University.

*Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60201. I thank Doug Bernheim and James Poterba for their comments.

consumption demand, labor supply equations, and the identity $K = F(K,l) - \delta K - c - g$ results in a pair of differential equations for K and λ which describe equilibrium. Individual optimality also requires that neither consumption nor capital ever vanish. These conditions determine the economy's dynamic equilibrium course for fixed initial capital stock and fiscal policies.

An intuitive feature of equilibria in this model is the tendency to converge to a steady state where consumption, labor supply, and output are constant if tax and spending policies are also constant. In particular, the economy returns to its original state after a temporary policy shock. My infinite-horizon approach can make the useful distinction between temporary, anticipated, and permanent shocks.

There are conditions under which this representative agent model also represents the equilibrium of a disaggregated model. In particular, if utility is isoelastic and labor is inelastically supplied (i.e., $u = c^{\gamma+1}/(\gamma+1)$), the individual Euler equations can be written in terms of consumption, yielding

$$\dot{c} = c\big(p - \big(r(1-\tau_K) + \delta\theta\big)/(1-\theta)\big)/\gamma.$$

If all individuals face the same marginal tax rate and investment return these individual conditions aggregate to a "consumption function,"

$$\dot{C} = C\big(p - \big(r(1-\tau_K) + \delta\theta\big)/(1-\theta)\big)/\gamma.$$

A similar aggregation with elastic labor supply is possible if $u = (c^{\gamma} + l^{\gamma+1})/(\gamma+1)$. For these cases, my model can examine "class" effects of various policies.

Before discussing the applications of this model, it is useful to indicate its advantages. First, it is a simple model in which movements of economic aggregates can be examined. Second, equilibria are locally unique, a distinct advantage over overlapping-generations models. Third, it is easy to examine the dynamics around a steady state, an approach to dynamics which is more realistic than the alternative of two-period general equilibrium models wherein the second period is also the last. Fourth, only a model with many periods can minimize intertemporal aggregation problems and be believably parameterized with estimates of technology and tastes.

Fifth, a study of the economy's dynamics around a steady state is useful in thinking about the incentives and tradeoffs which exist in a rational expectations policy equilibrium. Rational expectations imply that *in equilibrium* the government will *choose* to do what is expected; however, the government's alternatives, the paths not chosen, are *unexpected* policy shocks. Examining such shocks gives information about the incentives facing the government. Finally, it is easy to expand the model to incorporate extra elements of realism. One example of this flexibility will be my discussion of uncertain fiscal policy.

There are important elements which this model ignores for reasons for tractability. The infinite-life assumption implicitly makes strong assumptions about the nature of bequest behavior. Other analyses, such as Alan Auerbach, Laurence Kotlikoff, and Jonathan Skinner (1983), make the alternative extreme assumption that there is no bequest motive. For issues related to intergenerational distribution, that approach surely dominates. However, for aggregate issues it is unclear which is preferred. Ultimately the choice rests on empirical determination of the nature of bequest motives. Until then it is valuable to examine a variety of paradigms since such a study will indicate which issues are robust to these specifications.

II. Fiscal Policy

I first discuss the critical positive features of the economy's response to current and anticipated tax changes. In all exercises below, it is assumed that the economy has converged to the steady state corresponding to a previous constant spending and tax policy, and that the policy shock eventually settles down. For example, suppose that τ_K is currently .5, that individuals had expected it to be .5 forever, and that we are currently in the corresponding steady state. One possible shock would be the announcement that τ_K will remain at .5 for two more years, be increased to .52 for one year, but then move back down to .5 permanently. Such a change would be a partially anticipated temporary increase in capital income taxation. I focus on such one-period policy changes, since any policy change is a sum of such elementary changes.

When we speak of changing a policy, it is obvious that we cannot change just one policy variable. If a tax is increased then revenues will change, necessitating a change in some other tax or spending policy. In all of the exercises in this section I shall assume that lump sum taxes adjust to balance the government's budget. This choice is the best for analyzing the impact of any single policy parameter.

Some propositions always hold in my model, while others depend on actual parameter values. One advantage of a multiperiod model is that one can use empirical estimates of the critical parameters, which themselves implicitly assume a multiperiod model. Below I will assert propositions which are true for most empirical estimates of the underlying parameters. My papers (1985a, 1987) provide a more complete accounting.

When one calculates the impact of such policy changes, important results are found, some obvious, some less so. First, announcements of future τ_K increases will reduce current investment. The assumed rebate of revenues implies that there will be no *direct* income effect. The increase in the price of future goods increases current consumption and reduces investment. A negative income effect arises since the efficiency of the economy is reduced, but it only dampens the increase in consumption.

Anticipations of increased future government consumption will increase current investment and usually increase current labor supply. This is due to a pure income effect since the increased government extraction of output will reduce the utility derived from private consumption. Since I assumed separability between c and g, this income effect is the only direct effect. Immediate and temporary increases in g will crowd out investment, however, due to consumption smoothing by consumers.

Anticipated future wage taxation will usually increase current investment and labor supply: workers work today when wages are relatively high. This substitution effect dominates if utility is additively separable in c and l, and holds for most empirical estimates of utility functions. However, immediate wage tax increases reduce labor supply and investment.

Anticipated future increases in the ITC will reduce current investment because investors will wait for the subsidy. Conversely, immediate and temporary ITC increases have a substantial positive effect on current factor supply, since agents want to take advantage of the subsidy while it exists.

An important feature of the short-run responses to future policies is that different discount rates are used when calculating different aspects of equilibrium. When computing the impact on revenue, the net interest rate is the appropriate discount rate to apply to future policies. However, a higher discount rate is applied to the policy change when calculating the impact on current consumption or labor supply. (See my 1985a paper.) This fact will be important below when I examine balanced-budget exercises.

III. The Efficiency Cost of Taxation

One important use of this kind of model is to evaluate the relative efficacy of various tax policy changes. This section discusses the efficiency cost of taxation where by efficiency cost I mean the wealth equivalent of the loss in utility due to using a distortionary tax instead of a lump sum tax to raise a dollar in revenue. The results are intuitive given the reactions described above. In discussing the quantitative importance of various effects, it will be convenient to refer to a basic example. For the purposes of illustration in this essay, I use the example of a Cobb-Douglas production function with capital share of .30 and steady-state depreciation equal to .12 of net output. Tastes will be $u(c,l) = \ln c + l^{.2}$. The time unit is chosen to be that period over which utility is discounted by .01. It will be assumed that $\tau_K = .4$ and $\tau_L = .3$.

In the short run, capital income taxation is essentially a lump sum tax on existing capital. However, future capital income taxation will be distortionary since it reduces investment. In general, the efficiency cost of a tax increase rises rapidly as it is more anticipated.

In my example, a τ_K increase announced today to be implemented in four periods and last for one generates an efficiency cost of 24 cents per dollar of revenue whereas the efficiency cost of a one-period increase 20 periods hence is 84 cents. An immediate and permanent increase has an efficiency cost of 60 cents. From these figures we see that only immediate and temporary capital taxation is like lump sum taxation.

In the case of labor taxation, the result is usually reversed. An immediate labor tax increase generates only a price effect reducing current labor supply. However, we saw above that increases in future labor tax rates will increase current labor supply and investment, and increase efficiency in the presence of factor taxation. Therefore, the efficiency cost of a wage tax increase falls as it becomes more anticipated. In my example, the efficiency cost ranges from 12 cents for an immediate one-period increase to a loss of 6 cents for a perfectly anticipated increase, with a permanent increase losing 8 cents. Since the ratio between cost of unanticipated and anticipated wage taxation is 2:1, and since they both exceed 5 cents, the static estimate, I find that dynamic considerations are important even when evaluating wage taxation.

Raising revenues by reducing the ITC is very costly in this model. In fact, total revenues may even be reduced because of the substantial impact on capital formation. In my example, revenue falls for immediate ITC reductions that last for 15 periods or less and the welfare cost of a permanent ITC reduction is $4 per dollar of revenue. Intuitively, these results occur because the ITC is a more targeted form of tax incentive for investment than τ_K reductions, being primarily a subsidy of capital formation instead of relief for old capital. It is particularly interesting that all classes may want increases in the ITC. In my example and many others (see my 1981 paper), even a permanent ITC increase financed by wage taxation is Pareto-improving, since the tax cost for workers is more than offset by increased worker productivity and wages.

While the exact efficiency cost of revenue in this model is very sensitive to parameter choices, the ranking of various policies is surprisingly robust. A permanent wage tax increase is less distortionary than a permanent capital income tax increase for all parameterizations suggested by the empirical literature and reducing the ITC is far more costly than raising either factor income tax. While this model cannot yield precise estimates as to the magnitude of efficiency gains from tax changes, it has strong implications as to the appropriate direction.

The major purpose of taxation is the financing of public consumption. When taxes are distortionary, the efficiency cost of the necessary extra revenue should be considered when evaluating a potential project. The usual argument is that the benefits should exceed the direct costs of a project, the excess representing the efficiency cost of taxation. However, the large excess burdens noted above do not imply that the critical benefit-cost ratios should substantially exceed unity. The reason is that an increase in future government consumption increases the current supply of both factors, alleviating the tax distortions. The result is that the premium which must be borne due to distortionary taxation is not nearly as large as the marginal efficiency cost of taxation. In my benchmark case, the appropriate benefit-to-cost ratio to use is unity if permanent wage taxation is used to finance permanent expenditure and 1.28 if permanent capital taxation is used. There is also a large reduction, up to 20 percent in my example, in the critical benefit-cost ratio if the expenditure is delayed. In some cases, the stimulus to current factor supply of future expenditure is so strong that if we are to finance a project requiring a constant expenditure by making a permanent increase in τ_L, then the critical ratio is *less* than one. These considerations argue that projects requiring large immediate expenditures would be held to a substantially higher standard than those that involve a steady or deferred stream of expenditures.

These same considerations also predict biases in government factor usage. For example, a capital-intensive, but immediate and temporary, project will face a tougher benefit-cost criterion than a labor-intensive one, since the latter will increase labor supply whereas the former will crowd out investment. On the other hand, labor-intensive long-term projects will face a tougher criterion than capital-intensive ones since anticipated demand for capital services stimulates factor supplies, whereas anticipated demand for labor services reduces them.

While the neglect of many aspects of reality, in particular, uncertainty and asset heterogeneity, make these results largely suggestive, they do point and the importance of dynamic considerations in even the most elementary policy problems. Furthermore, the approach taken in my paper (1987) can be used in more realistic models.

IV. Balanced Budget Changes

Many of the exercises above assumed that the government's budget is balanced by changes in lump sum taxation. I next discuss the impact of policy changes when government consumption and distortionary taxes are altered to balance the budget, a more common experience.

One of the most studied exercises of this type is the temporary reduction of taxes, causing increased reliance on debt in the short run, but ultimately leading to increased taxation. Standard Keynesian arguments, based on the finite life of agents or capital market imperfections, assert that such a policy shock will increase current consumption because individuals regard the new bonds as part of wealth. However, my paper (1985d) showed that consumption can *fall* for two reasons: reducing τ_K will cause future goods to be cheaper, and shifting taxation to future capital income generates a negative income effect since a lump sum tax is replaced by a distortionary one. Current output and investment increase, and interest rates fall. Furthermore, the magnitude of this effect is of the same order as the Keynesian effects. Labor tax changes will generate opposing results since its dynamic reactions differ. However, for the case of a uniform income tax, the τ_K effects dominate. Hence, introducing *distortionary* taxation into this debate shows that the net result is an empirical matter and argues for Ricardian equivalence as an appropriate benchmark.

An alternative way to finance a current cut in taxes is a cut in future spending. If this is the anticipation of private agents, then the price effect of lower interest taxation is countered by the income effect generated by less government extraction. As we push the spending cut into the future, both the debt and the future necessary spending cut will grow at the net interest rate. However, the impact of future spending changes on current consumption is discounted at a much higher rate. Hence the net effect of the budget-balancing future spending cut on current consumption will fall as the spending cut is delayed. While the net effect is theoretically ambiguous, the income effect dominates for most reasonable parameterizations of tastes and lags, implying an increased consumption and crowding out of investment in the short run, causing interest rates to rise and output to fall. When utility is separable between labor and consumption, a fall in labor supply will also result, aggravating these contradictionary impacts. Even when there is no immediate contraction, some period of falling investment and output will precede the spending cut.

The major conclusion from these cases is that there is no sharp contemporaneous relation between deficits, spending, output, and interest rates. A critical determinant is the expectation of how the deficit will ultimately be financed.

V. Uncertain Future Fiscal Policy

An unrealistic assumption of my analysis so far is that agents know future policy with certainty. It is clear that there is much uncertainty in reality and private agents are often heard indicating that such uncertainty affects their current actions. The model I have examined is modified in my paper (1985c) to discuss the impact of uncertainty about future policy.

The first principle that arises is the existence of a magnification effect. More specifically, a mean-preserving increase in the uncertainty of a policy's timing will preserve the direction and increase the magnitude of the policy's effect on current private decisions whenever the uncertainty about the magnitude of the change is small. For example, an announcement that in five years the capital income tax will be raised for one year will reduce current investment. An announcement that the temporary increase will, with equal probability, occur either four or six years from now will cause an even greater fall in current investment. This is somewhat surprising, since it contradicts conventional precautionary saving arguments.

This magnification effect has some interesting implications for various fiscal policies. We saw above that a temporary substitution of debt for capital income taxes will reduce consumption in the short run. Since an anticipated future τ_K rise increases current consumption, uncertainty in the date of a budget-balancing increase will reduce that initial fall in consumption. If the debt is to be financed by future spending cuts, then we saw that consumption increases. Since future spending cuts increase current consumption, uncertainty as to when the cut comes will make the rise in current consumption even greater. Therefore, uncertain timing reduces the net balanced-budget effect in the case of debt financed by future capital taxation, but magnifies the net effect if future spending cuts are expected.

One of the causes of uncertainty is that opposing political factions champion alternative policies. The outcome is often uncertain, particularly since it may hinge on noneconomic developments. We can examine the impact of such political battles on resource allocation in my model. In particular, assume a tax cut sends the budget into deficit, and that one side wants to balance the budget by future spending cuts, whereas the other side prefers to increase τ_K. Suppose that this debate continues until one side yields, and that there is a constant probability in each period that a side yields. In this model, the immediate response to the outbreak of this conflict depends on its expected duration. If a quick resolution is expected, then the outcome, whether a spending cut or tax increase, will act to increase current consumer spending and dominates the price effect of the short-lived lower τ_K. On the other hand, if resolution is expected only in the distant future then the price effect of the τ_K cut dominates, increasing investment and output.

VI. Redistribution and Taxation

Factor taxation is determined not only by efficiency considerations, but also by judgments of who should bear the burden. It would appear that little can be said about what tax policies are desirable, since all will want someone else to pay the taxes which finance public spending. However, if we have only factor taxation, then it turns out that much can be said about long-run rational tax policy. In particular, in a large class of models there is unanimity that capital should bear no tax burden in the long run.

Suppose we have two classes that differ by tastes, wealth, and labor endowment. Suppose that we examine a richer class of utility functions allowing pure rates of time preference, ρ, to depend on past and current consumption. Such specifications generate a long-run saving supply function with finite elasticity, whereas our original specification had an infinite long-run savings elasticity. Furthermore, assume that for any constant interest tax rate, a steady state existed in which both groups hold finite wealth. My paper (1985b) showed that any convergent Pareto-efficient tax policy will put no tax on capital in the long run. The general principle is that there should be no taxation of stocks, only flows.

This is clearly not a reasonable prediction as to what tax policy will be implemented, because almost any Pareto-efficient program is dynamically inconsistent. Implementing a dynamically inconsistent policy is difficult, since policies are really chosen sequentially and promises of future policies are seldom enforceable. One way to measure the importance of this is to ask how high could τ_K be before all would agree that a permanent τ_K increase, with the revenues going to the workers, would be detrimental. Such a tax rate would be a Phelps-Pollak solution to the dynamic consistency problem. That tax rate is usually between .3 and .6, being .5 in our example. This indicates that the incentives to deviate from the optimum plan will be substantial and that the dynamic inconsistency problem can be particularly costly in the context of factor taxation. It also points to the importance of extending this model in the strategic directions necessary to examine the effects of and solutions to dynamic consistency problems.

VII. Conclusions

This paper has reviewed some aspects of the theory of dynamic factor taxation in a rational expectations, representative agent model of dynamic general equilibrium. Many important extensions are desirable, particularly the incorporation of uncertainty in production and rigorous modeling of the determination of policy choices in the political arena. Also, explicit incorporation of the market imperfections implicit in conventional Keynesian macroeconomic would enrich the model. However, I have shown that this model can be profitably used to examine the impact of dynamic taxation policy, anticipated and unanticipated, certain and uncertain, on resource allocation decisions, efficiency, and the distribution of wealth.

REFERENCES

Auerbach, Alan, Kotlikoff, Laurence and Skinner, Jonathan, "The Efficiency Gains from Dynamic Tax Reform," *International Economic Review*, February 1983, *24*, 81–100.

Brock, William and Turnovsky, Stephen, "The Analysis of Macroeconomic Policies in Perfect Foresight Equilibrium," *International Economic Review*, February 1981, *84*, 179–209.

Hall, Robert, "The Dynamic Effects of Fiscal Policy in an Economy with Foresight," *Review of Economic Studies*, April 1971, *38*, 229–44.

Judd, Kenneth L., "Dynamic Tax Theory: Exercises in Voodoo Economics," mimeo., 1981.

______, (1985a) "Short-Run Analysis of Fiscal Policy in a Simple Perfect Foresight Model," *Journal of Political Economy*, April 1985, *93*, 298–319.

______, (1985b) "Redistributive Taxation in a Simple Perfect Foresight Model," *Journal of Public Economics*, October 1985, *28*, 59–84.

______, (1985c) "The Macroeconomic Effects of Uncertain Fiscal Policy," mimeo., 1985.

______, (1985d) "Debt and Distortionary Taxation in a Simple Perfect Foresight Model," mimeo., 1985.

______, "The Welfare Costs of Factor Taxation," *Journal of Political Economy*, forthcoming 1987.

Evaluating Fiscal Policy with a Dynamic Simulation Model

By ALAN J. AUERBACH AND LAURENCE J. KOTLIKOFF*

Those schooled in the shifting curves of static and steady-state macro models may not fully appreciate the dynamic nature of fiscal policy. Simple blackboard models can convey neither the timing nor the magnitude of responses to short- and intermediate-term fiscal policies, nor can they isolate the impact of fiscal policies on transitional generations. There is also a range of issues, such as deficit finance and the relative efficiency of alternative tax structures, that cannot be properly addressed without solving for the economy's transition path.

Recent experience has provided several experiments in dynamic fiscal policy, including the accumulation of large amounts of official government debt, expected future changes in the level of social security benefits, shifts in the tax structure, and increases and then reductions in investment incentives. Each of these policies has important transitional as well as long-term effects. The analysis of these effects is possible using a dynamic general equilibrium numerical simulation model.

I. A Dynamic Simulation Model

In a series of our papers (1983a, b, c, d; 1985a, b; and our paper with Jonathan Skinner, 1983), and in a book (1987), we have developed such a dynamic general equilibrium numerical simulation model. We have used the model to study a range of fiscal policies, including deficit finance, changes in the tax structure, changes in the degree of tax progressivity, increases in investment incentives, and Social Security reform. We have also used the model to study the incidence and efficiency of alternative fiscal policies and to study the impact of demographics on saving and economic growth.

Like other numerical simulation models, this model provides an alternative to comparative statics analysis. The practice of signing derivatives is replaced here with simulations of policy changes and sensitivity analysis of how policy changes depend on taste and technology parameters. The results obtained with such a model are no less general than those obtained using comparative statics. Simulation models have the additional advantage of applying to large changes in tax policy instead of the infinitesimal changes for which the derivatives of comparative statics are valid. This ability to study large changes is very important in the analysis of economic efficiency; one of the most basic results in public finance is that the deadweight loss of a tax rises in proportion not to the tax rate itself, but to its square. Such nonlinearities may be especially important when one is considering the simultaneous application of several policies.

II. The Model

The simulation model is described in detail in our book. However, a broad outline of its characteristics and the method of solution will provide a basis for interpreting the results presented below.

The model is composed of households, firms, and the government. Households live for 55 periods (age 20 to 75). They are assumed to have rational expectations and to maximize a CES lifetime utility function of consumption and leisure subject to the budget constraint that the present value of consumption not exceed the present value of after-tax labor income plus transfers. There are also nonnegativity constraints on the labor supply of each individual in each cohort at each age. When the shadow wages associated with these constraints are positive, the individual is retired.

*University of Pennsylvania, Philadelphia, PA 19104, and Boston University and NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, respectively.

Our CES utility function has constant intertemporal elasticities of substitution, γ and ρ, respectively. It also has a time preference rate δ, and a leisure share parameter, α. Based on empirical evidence about these parameters, we set $\alpha = 1.5$, $\delta = .015$, $\rho = .8$ and $\gamma = .25$. In all simulations reported here, a 1 percent annual population growth rate is also assumed. The budget constraint depends not only on the interest rate r and the wage profile w_t (our cross-section age-wage profile is based on estimate of Finis Welch, 1979), but also on the average tax rates on capital income, labor income, and consumption, the payroll tax used to finance Social Security, and the level of Social Security benefits. In cases when the tax system is progressive, the average tax rates vary with the size of the tax base. This dependence is considered in the optimization decision, with both marginal and average tax rates affecting the household's choices.

The model has a single output, produced by identical firms that are assumed to behave competitively and to have a CES production function in labor and capital. The production function is normalized so that the wage in the base case is 1.0. In the base case, we use a Cobb-Douglas production function with capital's income share equal to .25.

Labor is assumed to be a variable factor of production, leading firms to set the marginal product of labor equal to the gross wage. Capital is assumed not to depreciate, and changes in the capital stock are subject to quadratic adjustment costs. This convex cost of adjustment leads to the smoothing of investment, so that outside of the steady state, the marginal product of capital will not necessarily equal the interest rate, and the value to the firm of an additional unit of capital may diverge from its replacement cost. The ratio of this market value to the replacement cost of capital is Tobin's "q" ratio (Fumio Hayashi, 1982).

There are actually two sources of variation of the value of the firm per unit of capital: changes in the value of marginal q associated with the costs of adjusting to new investment and changes in the relative valuation of old and new capital due to changes in the tax-based distinction between new and existing capital goods. For example, such a distinction arises when new capital receives investment incentives, such as the investment tax credit, for which existing assets do not qualify. Both adjustment costs and tax distinctions between new and old capital can produce significant changes in stock market values of capital.

The government in this model has a main fiscal authority that raises taxes to pay for government spending on goods and services and a separate social security system which, like the actual U.S. Social Security system, is unfunded. The government's budget constraint is that the present value of taxes equals the present value of government spending plus the initial stock of debt.

III. The Model's Solution

The calculation of the equilibrium path of the economy given a particular parameterization typically proceeds in three stages: 1) solving for the long-run steady state of the economy before the assumed change in fiscal policy begins, 2) solving for the long-run steady state to which the economy eventually converges after the policy takes effect, and 3) solving for the transition path the economy takes between these two steady states. The perfect foresight assumption is critical only in this third state, since any model of expectations formation would predict correct expectations in a steady state. The solution method is one of iterating to find a fixed point.

After solutions for the initial and final steady states of the economy are found, the economy's transition path is calculated by 1) providing the economy with 150 years to reach the new steady state (many more years than it actually takes for the model to reach a position indistinguishably close to the final steady state), and 2) solving for behavior during those 150 transition years fixing expectations for years after 150 at the final steady-state values that will, in fact, obtain. Variations in initial guesses and the number of years permitted for the transition to take place have never produced changes in the solutions obtained.[1]

IV. Simulation Results

Some selected simulations provide an indication of the surprising aspects of fiscal policy not evident from earlier static analyses. We focus on three issues that have been of particular interest to economists in recent years: the choice of tax base, the impact of investment incentives, and the effects of budget deficits.

A. *Income vs. Consumption Taxes*

Since the careful studies produced by the U.S. Treasury (1977) in the United States and the Meade Committee (Institute for Fiscal Studies, 1978) in the United Kingdom, and the influential papers by Martin Feldstein (1978) and Michael Boskin (1978), economists have wondered whether reducing or removing capital income taxation's discrimination against future consumption would increase economic efficiency. Since time invariant proportional taxation of consumption does not affect the rate of return available to saving, the discussion has often focused on switching from the income tax to a personal consumption tax rather than simply removing capital income from the tax base (see, generally, Joseph Pechman, 1980).

The basic efficiency argument is that the increased labor supply distortion of such a switch would be more than offset by the reduced saving distortion. As is well-known, this is a question of "second-best" economics to which there is no simple general answer, but the academic literature including papers mentioned above and, in particular, one by Lawrence Summers (1981), argued passionately that the efficiency gains from reducing capital income taxation would be quite large because of the relatively high elasticity of savings with respect to the interest rate.

Static efficiency calculations ignore what is probably the most important issue in the switch from income to consumption taxation: the intergenerational redistribution of the tax burden. Since consumption tends to occur later in life than income, a switch to consumption taxation shifts each year's tax burden toward the elderly. The result is that the current elderly population pays more, while subsequent generations pay less by having their tax payments deferred to older age. This provides a substantial increase in the long-run utility of generations in the eventual steady state, equivalent, in our model, to a permanent increase in consumption and leisure of about 6 percent assuming an initial income tax rate of 30 percent.

Removing capital income taxation directly from the proportional income tax base, that is, switching to a wage tax, while equivalent in a static model to adopting a proportional consumption tax, has quite different results in a dynamic model, since there is an opposite tax windfall. The existing elderly population gains from the shift in tax burden to wages, which occur earlier in life on average than income. This makes them better off, but makes all subsequent generations worse off; in the eventual wage-tax steady state welfare is reduced by over 4 percent of lifetime consumption and leisure. The very different intergenerational transfer effects of these two "equivalent" tax policies are shown in Figure 1.

[1] Previous analyses of uniqueness with overlapping generations models (for example, Guillermo Calvo, 1978, and Timothy Kehoe and David Levine, 1985) have provided examples in which there is a continuum of transition paths to the new equilibrium. The nonuniqueness problem arises if there are more stable roots to the linearized version of the system in the neighborhood of the final steady-state equilibrium than initial conditions. The requirement of convergence eliminates only the unstable roots, leaving, in some cases, a continuum of feasible paths that satisfy the initial conditions. While we have not explicitly calculated the roots of a linearized version of our model, such analysis has been conducted for a similar model by John Laitner (1984). He found the transition path to be determinate, with the number of stable roots equal to the number of initial conditions. His results, together with our own findings that, in practice the solution calculated by our model does not depend on the initial guesses chosen for the transition path, strongly suggest to us that indeterminacy is not a problem.

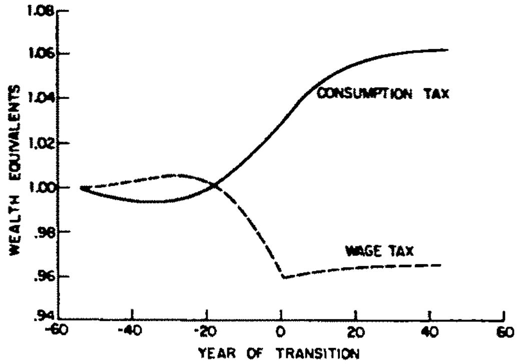


FIGURE 1. WEALTH EQUIVALENTS

Thus a dynamic analysis shows that the long-run impacts of switching to consumption vs. wage taxation are quite different. So too are the efficiency impacts of the switches to consumption and wage taxation. By taxing the consumption financed by preexisting wealth, the consumption tax base equals the initial stock of wealth plus the present value of all future wages, rather than just the latter as under a wage tax (Christophe Chamley, 1983). This tax on initial wealth arising under the consumption tax is a lump sum tax and explains why the consumption tax is more efficient than the wage tax.

To analyze the efficiency gains of switching tax bases our model includes a Lump Sum Redistribution Authority (LSRA) that transfers resources across generations in a lump sum fashion. In our efficiency transition calculations, the LSRA maintains the preexisting utility levels of generations initially alive at the time of the tax change, and any efficiency gains (losses) are allocated across subsequent generations in such a way that all subsequent generations enjoy a uniform increase (decrease) in utility. In these LSRA transitions, we found that switching from the 30 percent income tax to the equal revenue consumption tax permits an increase in utility for all subsequent generations which is equivalent in the initial steady state to a permanent increase in lifetime consumption and leisure of 1.7 percent, while abolition of capital income taxes, that is, switching to wage taxation, induces a *decline* of 2.3 percent. Hence, perhaps 60 percent of the difference between the non-LSRA changes in long-run welfare under labor income taxation and consumption taxation is attributable to intergenerational transfers.

It should be emphasized that certain policies that appear to resemble the consumption tax, such as expanding the limits on contributions to individual retirement accounts, do not offer the efficiency gains of consumption taxation because they share with the repeal of capital income taxes the crucial feature of exempting from taxation the consumption of existing wealth.

The presence of progressivity reinforces these findings for two general reasons. First, since distortions are worse with higher marginal tax rates, any efficiency gains associated with a reduction in distortions will be magnified. Second, the fact that average as well as marginal tax rates under progressive taxation rise with the tax base reinforces the shift in the tax burden and its associated distortions to the elderly under a consumption tax and toward the working population under a labor income tax. Hence, the distinction between these two "equivalent" tax bases is even greater than indicated by the proportional tax simulations.

B. *Investment Incentives*

The increase in accelerated depreciation allowances under the Economic Recovery Tax Act of 1981 was viewed by many as a windfall to corporations and the owners of corporate shares. More careful analysis suggests the opposite, and this is confirmed by our simulation results. First, consider the theoretical impact of investment incentives. The introduction or enhancement of investment incentives not only encourages investment, it also lowers the present value of taxes on new investment, while leaving unchanged the present value of taxes on old capital. Because old capital is at a tax disadvantage, its market value must fall. In the case of an investment tax credit, for example, the effect will be to drive the value of old capital down to the cost of new capital net of the investment tax credit, for which only new capital qualifies. The short-run impact of adjustment costs, on the other hand, will be to mitigate this fall in the stock market value of old capital; with adjustment costs, old capital earns rents on the installation of new capital.

A drop in the value of capital, combined with a cut in the tax burden on new investment, is good for savings but bad for old people, just like a consumption tax. In fact, investment incentives in the presence of an income tax are not only just like consumption taxes; they are consumption taxes. This equivalence is most easily seen for the ultimate acceleration of depreciation allowances, immediate expensing, though it is just as true in other cases. Under expensing, the firm pays taxes on income net of new investment, just as an individual would under a consumption tax, since income net of saving equals consumption. And, as under a consumption tax, all income net of saving is taxed, even though some of it may be income earned on preexisting assets. The only difference is that in this case the tax is collected at the firm level and capitalized into the value of existing capital goods, rather than paid by the individual upon consumption. It thus represents a firm-level rather than an individual-level consumption tax.

Given this equivalence, it is somewhat surprising that a policy that in one form is seen as so unfair to the owners of capital is seen when presented in a different form as so unfair to everyone else. The fact that adjustment costs may offset the windfall loss to existing capital caused by the introduction of investment incentives does not change the equivalence. It would be present as an absolute increase in the value of assets under the direct consumption tax.

Our simulations suggest that the windfalls associated with a move to investment expensing may be quite large. For an adjustment cost parameter of $b=10$ (on the low end of empirical estimates, but by no means small), we find that a move from a 15 percent income tax to the same tax with complete expensing (i.e., a consumption tax) reduces the value of the existing capital stock by nearly two-thirds the size of the tax rate cut on new investment, or about 9.5 percent.

There are other results that may appear surprising but become less so when the equivalence of investment incentives and consumption taxes is remembered. For example, it is quite possible for an increase in investment incentives to raise revenue without the economy being on the wrong side of the Laffer curve. In our paper (1983b), we present a simulation in which investment incentives are introduced and financed, in the short run, by issuing debt. In the simulation the income tax rate is held constant at 30 percent for a period of 20 years. During this time the increase in the tax base associated with the investment incentives leads to an increase in revenue which is sufficient to retire the government debt issued at the beginning of the transition. After year 20, the income tax rate must be reduced to prevent accumulating an infinite surplus. Thus, the crowding in from switching towards a consumption tax base exceeds the crowding out from the short-run deficits. While the investment incentives are cuts in business taxes, not everyone in the model experiences a tax cut, and the distributional effects of the shift in the tax burden lead to the observed outcome.

C. *Deficits, Crowding Out, and Crowding In*

As recent experience suggests, most students of fiscal policies appear to believe that deficits arising from tax cuts will be associated with short- and long-run crowding out of capital and short- and long-run increases in interest rates. While our simulation studies of deficit policies confirm these long-run predictions, we find that, except for tax cuts of very long duration, deficits arising from tax cuts will be associated with short-run *crowding in* of capital and short-run *declines* in interest rates. The simple explanation is that tax cuts have substitution as well as income effects. In the short run, individuals take advantage of the temporarily low tax rate on wages and the return to capital by working and saving more. One response to this line of argument is that substitution elasticities are potentially small. While this may be true, our model assumes fairly small substitution elasticities. What is not, however, typically understood is that although substitution elasticities are small, the change

TABLE 1—ECONOMIC IMPACT OF DEFICIT FINANCING

	S/Y	W	r	L	K
Crowding Out under Alternative Deficit Policies					
Initial Steady State	0.037	1.000	0.067	19.10	95.1
1-Year Income Tax Reduction					
Year:					
1	0.049	0.991	0.069	19.80	95.1
2	0.034	1.002	0.067	19.04	95.4
3	0.034	1.001	0.067	19.04	95.3
4	0.035	1.001	0.067	19.04	95.3
5	0.035	1.001	0.067	19.04	95.3
10	0.036	1.000	0.067	19.05	94.9
30	0.037[a]	0.998	0.067	19.07	94.2
60	0.037[a]	0.997	0.068	19.08	93.9
90	0.037[a]	0.997	0.068	19.08	93.9
Final Steady State	0.037[a]	0.997	0.068	19.08	93.9
5-Year Income Tax Reduction					
Year:					
1	0.046	0.992	0.069	19.76	95.1
2	0.045	0.992	0.068	19.74	95.3
3	0.044	0.993	0.068	19.73	95.6
4	0.043	0.994	0.068	19.71	95.8
5	0.026	1.006	0.066	18.83	95.9
10	0.028	1.002	0.067	18.87	94.6
30	0.032	0.992	0.069	18.96	91.3
60	0.036	0.987	0.070	19.02	89.9
90	0.036	0.987	0.070	19.02	89.7
Final Steady State	0.036	0.987	0.070	19.02	89.7
20-Year Income Tax Reduction					
Year:					
1	0.034	0.994	0.068	19.58	95.1
2	0.033	0.994	0.068	19.56	95.0
3	0.031	0.994	0.068	19.55	94.9
4	0.030	0.994	0.068	19.53	94.8
5	0.029	0.993	0.068	19.51	94.6
10	0.023	0.991	0.068	19.45	93.3
30	−0.014	0.964	0.075	17.72	76.1
60	0.011	0.888	0.096	18.08	56.0
90	0.020	0.867	0.103	18.11	50.8
Final Steady State	0.023	0.856	0.107	18.13	48.5

Note: S = net national saving; Y = net national product; W = wage rate; r = interest rate; L = aggregate labor supply; K = capital stock.

[a] This saving rate is below that in the initial steady state to the fourth decimal.

in tax rates may be substantial and, therefore, have a substantial impact on relative prices of consumption vs. leisure, and consumption today vs. consumption tomorrow.

Table 1 reports the simulation results of 33 percent cuts in the income tax rate lasting 1 year, 5 years, and 20 years. The table gives the saving rate, S/Y, the income tax rate, W, the pretax interest rate r, the level of labor supply, L, and the stock of capital, K. Note that in the 1- and 15-year tax cuts the saving rate is larger in year 1 and in years 1 to 14, respectively, than in the initial steady state; hence, crowding in occurs under these short-term tax cut policies until the tax rate is increased. After the tax rate is increased crowding out proceeds, but fairly slowly.

The example of short-run crowding in arising from short-run tax cuts demonstrates that a policy that is ultimately detrimental to capital formation can appear, in the short run, to be increasing savings. Thus there is the potential to misread policy by focusing too strongly on the short-run impacts.

V. Conclusion

Dynamic simulation models can resolve a number of important issues that cannot be adequately considered in static or steady-state analyses. The results from dynamic analyses can be quite surprising. We, at least, were surprised to learn that deficits are most likely to cause short-run crowding in and lower short-term interest rates, that investment incentives are detrimental to capitalists, that business tax cuts may be self-financing, that the degree of tax progressivity is as important as the choice of tax base for issues of savings and efficiency, that official government debt is a highly unreliable measure of the government's true economic debt policy, and that while baby "busts" like those underway in the United States are bad for social security, they are, on net, likely to be beneficial to the economy.

REFERENCES

Auerbach, Alan J. and Kotlikoff, Laurence J., (1983a) "National Savings, Economic Welfare, and the Structure of Taxation," in Martin Feldstein, ed., *Behavioral Simulation Methods in Tax Policy Analysis*, Chicago: University of Chicago Press, 1983.

______ **and** ______, (1983b) "Investment Versus Savings Incentives: The Size of the Bang for the Buck and the Potential for Self-Financing Business Tax Cuts," in L. H. Meyer, ed., *The Economic Conse-*

quences of Government Deficits, Boston: Kluwer-Nijhoff, 1983.

______ **and** ______, (1983c) "Social Security and the Economics of the Demographic Transition," in H. Aaron and G. Burtless, eds., *Retirement and Economic Behavior*, Washington: The Brookings Institution, 1983.

______ **and** ______, (1983d) "An Examination of Empirical Tests of Social Security and Savings," in Elhanan Helpman et al., eds., *Social Policy Evaluation: An Economic Perspective*, New York: Academic Press, 1983.

______ **and** ______, (1985a) "Simulating Alternative Social Security Responses to the Demographic Transition," *National Tax Journal*, June 1985, *38*, 153–68.

______ **and** ______, (1985b) "The Efficiency Gains from Social Security Benefit-Tax Linkage," NBER Working Paper No. 1645, June 1985.

______ **and** ______, *Dynamic Fiscal Policy*, Cambridge: Cambridge University Press, 1987.

______, ______, **and Skinner, Jonathan,** "The Efficiency Gains from Dynamic Tax Reform," *International Economic Review*, February 1983, *24*, 81–100.

Boskin, Michael J., "Taxation, Saving, and the Rate of Interest," *Journal of Political Economy*, April 1978, *86*, S3–27.

Calvo, Guillermo, "On the Indeterminacy of Interest Rates and Wages with Perfect Foresight," *Journal of Economic Theory*, December 1978, *19*, 32–37.

Chamley, Christophe, "Efficient Tax Reform in a Dynamic Model of General Equilibrium," mimeo., World Bank, 1983.

Feldstein, Martin, "The Welfare Cost of Capital Income Taxation," *Journal of Political Economy*, April 1978, *86*, S29–51.

Hayashi, Fumio, "Tobin's Marginal and Average *q*: A Neoclassical Interpretation," *Econometrica*, January 1982, *50*, 213–24.

Kehoe, Timothy J. and Levine, David, "Comparative Statics and Perfect Foresight in Infinite Horizon Economies," *Econometrica*, March 1985, *53*, 433–53.

Laitner, John, "Transition Time Paths for Overlapping-Generations Models," *Journal of Economic Dynamics and Control*, May 1984, 7.

Pechman, Joseph A., *What Should Be Taxed: Income or Expenditure?*, Washington: The Brookings Institution, 1980.

Summers, Lawrence H., "Capital Taxation and Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, *71*, 533–44.

Welch, Finis, "Effects of Cohort Size on Earnings: The Baby Boom Babies' Financial Bust," *Journal of Political Economy*, October 1979, *87*, S65–97.

Institute for Fiscal Studies, *The Structure and Reform of Direct Taxation*, Report of a Committee Chaired by Professor J. E Meade, London: Allen and Unwin, 1978.

U.S. Treasury, *Blueprints for Basic Tax Reform*, Washington: USGPO, 1977.

Growth Based on Increasing Returns Due to Specialization

By PAUL M. ROMER*

This note describes an attempt to model increasing returns that arise because of specialization. The idea that increasing returns and specialization are closely related is quite old, but, apparently for technical reasons, we have no fully worked out dynamic model of growth along these lines. There are now several models of growth that consider increasing returns that arise from the accumulation of knowledge. (See, for example, my dissertation, 1983, and 1986a paper; Robert Lucas, 1985; Edward Prescott and John Boyd, 1987.) Despite the presence of aggregate increasing returns, these models can support a decentralized competitive equilibrium with externalities; the externalities arise because of spillovers of knowledge. At least since the publication of Kenneth Arrow's 1962 paper on learning by doing, it has been clear that a competitive equilibrium with externalities provides a tractable framework for the study of increasing returns in a dynamic model. The model described here shows that a closely related framework can be used to study specialization.

The idea that specialization could lead to increasing returns is as old as economics as a discipline. The idea that a decentralized equilibrium with externalities could exist despite the presence of aggregate increasing returns is as old as the notion of an externality. In *Principles of Economics*, Alfred Marshall introduces the notion of an "external economy" to justify the use of a decentralized, price-taking equilibrium in the presence of aggregate increasing returns. He notes in passing that an increase in "trade-knowledge" that cannot be kept secret represents a form of external economy (p. 237). He gives more emphasis to the growth of subsidiary trades that use "machinery of the most highly specialized character" (p. 225), claiming that these too give rise to some vague sort of external effect. In the spirit of specialized endeavors, the model presented below ignores increasing returns from investments in knowledge and external effects due to spillovers of knowledge. It focuses exclusively on the role of specialization. A more realistic and more ambitious model would examine both effects.

I. Static Models of Specialization

The first step in the construction of a model where specialization leads to a form of increasing returns has been taken by Wilfred Ethier (1982). He suggests that we reinterpret as a production function the utility function used by Avinash Dixit and Joseph Stiglitz (1977) to capture a preference for variety. In this reinterpretation, the output of final consumption goods is an increasing function of the total number of specialized intermediate inputs used by a final goods producer. In a continuum version of this model, the list of intermediate inputs used in final good production is a function $x\colon \mathbb{R}_+ \to \mathbb{R}$, where $x(i)$ denotes the amount of intermediate good i used. A production function using both labor and intermediate inputs that is analogous to the Dixit-Stiglitz utility function is

$$(1)\qquad Y(L,x) = L\int_{\mathbb{R}_+} g\left(\frac{x(i)}{L}\right) di,$$

where g is an increasing, strictly concave function with $g(0)=0$. In the special case considered by Dixit-Stiglitz and by Ethier, g is the power function $g(x)=x^\alpha$, with $0<\alpha$

[*] Department of Economics, University of Rochester, Rochester NY, 14627.

<1. Then Y takes on the more familiar form

$$(2)\qquad Y(x)=L^{1-\alpha}\int_{\mathbb{R}_+} x(i)^{\alpha}\,di.$$

Let $\{N,M\}$ denote the list of inputs $x(i)$ that takes on the constant value $x(i)=N/M$ on the range $i\in[0,M]$. Thus, M measures the range or number of intermediate inputs used, and N measures the total quantity of such inputs. The graph of $x(i)$ is a rectangle of width M lying on the i axis and having a total area equal to N. In general,

$$(3)\quad Y(L,\{N,M\})=LMg(N/LM).$$

If g is a power function, this becomes

$$(4)\quad Y(L,\{M,N\})=M^{1-\alpha}(L^{1-\alpha}N^{\alpha}).$$

In either case, it is easy to show that output of the final good increases with M, the range or number of different inputs, when labor and the total quantity of intermediate inputs are held constant. This loosely captures the idea that a *ceteris paribus* increase in the degree of specialization increases output. In equation (4), Y appears to exhibit increasing returns to scale, but N and M are not the relevant inputs. As a function of labor L and the lists of intermediate inputs $x(i)$, Y is a concave production function that is homogeneous of degree 1.

To capture the idea that fixed costs limit the degree of specialization, assume that the intermediate inputs $x(i)$ are produced from a primary input Z according to a cost function that has a U-shaped average cost curve. Preserving the symmetry in the model, assume that an amount $x(i)$ of any good i can be produced at a cost $h(x(i))$. Inaction at zero cost is feasible, so $h(0)$ equals zero; but at any positive level of production, $h(x)$ is greater than some quasi-fixed cost $\bar{h}$. For simplicity, I assume that this cost is measured purely in terms of the primary input and ignore labor inputs in the production of intermediate inputs. Since this cost is measured in units of the primary good per unit of infinitesimal length di, the resource constraint faced by the economy as a whole is

$$(5)\qquad \int_{\mathbb{R}_+} h(x(i))\,di\le Z.$$

With this specification for costs, the feasible range of intermediate inputs is finite.

Together, a production function like Y and a cost function like h offer an extremely crude representation of the many specialized goods that are in fact used in multiple stages of production. It is intended only as a kind of reduced form. (See Spyros Vassilakis, 1986, for an alternative, more detailed model of specialization.) Modeling the output of a firm in the consumption goods sector as a deterministic function of the entire set of available specialized inputs is a convenient simplification that cannot be taken literally. Besides allowing for multiple stages of intermediate inputs, a more realistic approach would extend this model in precisely the way that Michael Sattinger (1984), Jeffrey Perloff and Steven Salop (1985), and Oliver Hart (1985) extend the Dixit-Stiglitz model of consumer preferences, allowing for many producers of final goods, each of whom has a technology that is most productive with a specific, small subset of all potential intermediate inputs. If the particular inputs that are most productive are distributed symmetrically across a large number of firms producing the final good, the aggregate effect should be similar to that achieved in the model here. If one allows for the possibility of household production, the model can accommodate an apparent preference for variety on the part of consumers as well. (Kenneth Judd, 1985, Nancy Stokey, 1986, and James Schmitz, 1986, are examples of dynamic models with preferences similar to the production function used here.) Ski boots and screw drivers have as much claim to be called intermediate inputs as pig iron and petrochemicals.

A decentralized equilibrium for this economy consists of a continuum of firms in the intermediate goods sector and an indeterminant number of firms producing final output goods with the constant returns to scale production function Y. The final goods firms are assumed to be price takers in all of their markets. Each of the intermediate input producing firms is the single producer of a particular intermediate input and has power in the market for its specialized good. It is still a price taker in the market for the primary input Z. Using final output goods as numeraire, let R denote the price of a unit of the resource Z. (The notation R will more appropriate in the next section where Z is a durable stock in a dynamic model and R has the interpretation of a rental rate.) Assuming for simplicity that the primary input has no alternative use in consumption or production, preferences can be any increasing function of final good consumption. For now, all that I need to specify about the demand side of the economy is that the individual consumers are price takers, and that they are endowed with the stock of the primary resource and an inelastically supplied quantity of labor.

The kind of equilibrium that obtains is a monopolistically competitive equilibrium similar to the one described by Dixit and Stiglitz. Given a list of prices $p(i)$ for the intermediate inputs that are produced, it is straightforward to derive demands for these inputs. Setting the aggregate supply of labor L equal to 1, the (inverse) demand function for any particular input i is proportional to the derivative of the function g that appears as the integrand in Y:

$$(6) \qquad p(i)=g'(x(i)).$$

Potential and actual producers of intermediate goods maximize profits taking these demand curves and the price R for the primary resource as given. (My 1986b paper describes a sequence of finite economies that rationalize this as a limit equilibrium.) In equilibrium, some goods i are produced, others are not. All firms in the intermediate goods industry (both potential producers and actual producers) earn zero profits. Given the derived demand curves, profit maximization on the part of intermediate goods producers leads to values of $x(i)$ that depend on the price of the primary resource R. The price R is determined by the requirement that profits for the intermediate goods producers must be zero.

For given Z, the key quantities to be determined are M, the number or range of intermediate inputs that are produced, and $\bar{x}$, the amount of each of these inputs that is produced. By the symmetry in the model, it is clear that all goods that are produced will be produced at the same level. To illustrate the equilibrium in a particular case, let g be the power function described above, and let the cost function h take the form $h(x)=(1+x^2)/2$. Then the equilibrium quantities are

$$(7) \qquad x(i)=\bar{x}=(\alpha/(2-\alpha))^{1/2},$$

on a set of inputs i of length

$$(8) \qquad M=Z(2-\alpha),$$

with $x(i)=0$ otherwise. The equilibrium value of R can be explicitly calculated, but is not revealing.

It is also straightforward to calculate the quantities that would be chosen by a social planner who maximizes output subject to the constraints imposed by the technology. A curious feature of the choice of g as a power function is that the quantities from the first-best social optimum coincide with those in the decentralized equilibrium. This result relies crucially on the fact that the stock of Z is given. Explicit calculation shows that in the equilibrium, the marginal value of an additional unit of the resource Z is R/α, strictly bigger than the market price, R. In any extension of this model that allows an alternative use for Z, the decentralized equilibrium will differ from the first-best social optimum. In particular, any model that explains growth by allowing individuals to forego current consumption and accumulate additional units of the resource Z will necessarily have an equilibrium with less accumulation of Z than would be socially optimal.

Even with a given quantity of the primary resource Z, a different choice of the function g can lead to equilibrium values for $\bar{x}$ and M that differ from the values that would be chosen by a social planner. The suboptimality arises for two distinct reasons. The downward-sloping demand curve faced by actual producers of intermediate goods causes the equilibrium level of $\bar{x}$ to be too small (and therefore causes M to be too big.) An opposing effect arises because the introduction of a new intermediate input creates surplus for the producers of final goods that cannot be captured by the firm selling the input. New intermediate inputs are introduced up to the point where total costs equal payments to a firm producing an intermediate input, but under standard monopoly pricing these payments are smaller than the surplus created by the additional inputs. This effect causes M to be too small (and therefore causes $\bar{x}$ to be too big.) The case where the function g is a power function happens to be such that these two effects on the quantities $\bar{x}$ and M exactly cancel. However, both effects cause Z to be undervalued.

To highlight the divergence between the private and social gains from the introduction of new goods, it is useful to consider an example that removes the usual distortion arising from a divergence between price and marginal cost. To preserve the result that final output depends nontrivially on the range of inputs used, the function g must have some degree of curvature. Since the derived demand curve for an intermediate input curve is proportional to the derivative of g, this implies that demand must be downward sloping in some region. To insure that price equals marginal cost, the intermediate goods producer must face a demand curve that is horizontal in the relevant region.

Thus, suppose that the function g is at least twice continuously differentiable with the following properties. On the interval $[0, x_0]$, g is strictly concave, with $g(0)=0$, $g'(x_0)=1$. On the interval $[x_0, \infty)$, let g have a constant slope equal to 1. In the graph of g, let G denote the intercept that is defined by tracing the constant slope of 1 back to the vertical axis. Thus, for $x > x_0$, $g(x) = G + x$. The curvature in the interval $[0, x_0]$ is needed simply to satisfy the requirement that $g(0) = 0$ without violating continuity. The derived inverse-demand curve $p(i) = g'(x(i))$ is a differentiable curve that may or may not have a finite intercept. It is downward sloping on the interval $[0, x_0]$, and takes on the constant value of 1 on $[x_0, \infty)$.

Consider the output from $Y(L, x)$ with this functional form for g. As before, let $\{N, M\}$ denote the rectangular list of inputs with a range of M different specialized inputs each supplied at the level $x(i) = N/M$. If N/M is greater than x_0 (and by choice of a small enough x_0, this will be true for all relevant lists of inputs), the expression for output as a function of N and M is

$$Y(L, \{N, M\}) = GLM + N. \tag{9}$$

As before, this is increasing in the range of inputs M when total labor L and the total quantity of intermediate inputs N are held constant. With this function and the previous choice of the cost function $h(x) = (1 + x^2)/2$, it is easy to verify the following equilibrium quantities. (As above, set the total quantity of labor equal to 1.) First, guess that the equilibrium price R for the resource Z is equal to 1. Then the marginal cost of additional units of $x(i)$ measured in units of output goods is $Rh'(x) = x$. The assumption that x_0 is small relative to 1 then implies that marginal cost intersects the marginal revenue schedule at the point $(p, x) = (1,1)$, which lies in the range where the demand curve is flat; hence, marginal revenue coincides with the demand curve at this point. Since the price R for the primary resource is equal to 1, this is also a point on the average cost curve—in fact, it is the point of minimum average cost—so this corresponds to a potential equilibrium. Given that x_0 is small and provided that the demand price $g'(x)$ does not go to ∞ too rapidly as x goes to zero, the U-shaped average cost curve will lie above the demand curve for all other values of x, tangent only at the point $(1,1)$. If so, this will be the unique monopolistically competitive equilibrium. In this case, the equilibrium list of inputs $x(i)$ takes on the value 1 for a set of inputs i of measure $M = Z$ and is zero elsewhere.

It is also a simple matter to calculate the solutions to the social planning problem for this economy. For this form of the function g, the decentralized equilibrium leads to a range of output goods that is too small relative to that achieved in the first best social optimum. All firms that produce intermediate goods do so up to the point at which the marginal cost equals the marginal product, so there is no force to offset the tendency for the equilibrium to provide too small a range of inputs. Equilibrium output is $Y = Z(G+1)$, but the price of Z is $R=1$. For this form of the function g as well as for the previous one, the marginal product of Z is greater than its equilibrium price.

II. A Dynamic Model

One simple way to make the static model into a growth model is to allow for the accumulation of the primary resource Z, which is now interpreted as a durable, general purpose capital good. For simplicity, I treat the supply of labor as being exogenous and neglect both a labor-leisure tradeoff and population growth. The specification of intertemporal preferences is conventional,

$$(10) \qquad \int_0^{\infty} U(c(t))e^{-\rho t}\,dt.$$

In the examples that follow, I will assume that the utility function $U(c)$ take the isoelastic form

$$(11) \quad U(c) = (c^{1-\sigma}-1)/(1-\sigma), \qquad \sigma \in (0,\infty).$$

For convenience, let there be a continuum of identical consumers indexed on the interval $[0,1]$, each endowed with an amount $Z(0)$ of the initial stock of general purpose capital. So that I can work interchangeably with per capita and per firm quantities, let there be a continuum of firms in the final goods producing sector, also indexed on $[0,1]$, all producing at the same level. (Because of the constant returns to scale in this sector, this is harmless.) Consumers will rent their capital (i.e. their stock of Z) to intermediate goods-producing firms. These firms use it to produce intermediate inputs $x(i,t)$ according to the technology defined by the cost function h, so that the feasible set of intermediate inputs at every point in time is constrained by equation (5). The intermediate inputs can be interpreted either as a flow of nondurable goods produced by the general purpose capital devoted to the production of inputs of type i, or as a service flow from a durable, specialized capital good of type i, that is created by transforming general purpose capital into specialized capital.

Assuming once again that the aggregate supply of labor is equal to 1, each individual in this economy receives per capita output (equal to per firm output) of $Y(1,x)$. This must be allocated between consumption $c(t)$ and investment in additional capital Z. The simplest investment technology is one that neglects depreciation and permits foregone output to be converted one-for-one into new capital. Thus, assume that

$$(12) \qquad \dot{Z} = Y(1,x) - c.$$

Without considering the general problem of how to calculate a dynamic equilibrium with monopolistic competition for this model, it is possible to describe equilibria for the specially chosen functional forms considered here. (For a discussion of general methods for calculating equilibria of this type, see my 1986b paper.) Consider first the case described above where $g(x)$ has a slope of 1 for values of x greater than x_0. From the calculation of the static equilibrium with these functional forms, it is clear that the rental rate R (and now it is a true rental rate) on a unit of Z is equal to one unit of consumption goods per unit time. Since one unit of consumption goods can be converted into one unit of capital Z, the price of capital goods in terms of consumption goods must also equal 1. Thus the instantaneous, continuously compounded rate of return on investments in capital goods is 100 percent per unit time. This preserves the values calculated from the static model and makes sense if the unit used to measure time is roughly a decade. The discount rate ρ must also be scaled up to reflect this choice of time units. However, to ensure that growth will take place, the discount rate is assumed to be less than the return to savings; that is, ρ is assumed to be less than 100 percent.

The value for $\bar{x}$ is 1 and the range of goods $M(t)$ is equal to $Z(t)$. Hence, $N(t) = M(t)\bar{x} = Z(t)$. Since output is given by $Y(L, x) = GLM + N$ and L is assumed to take on the constant value 1, output at time t is $Y(t) = Z(t)(G+1)$. For the specified form of preferences, the instantaneous, continuously compounded interest rate on consumption good loans is $\rho + \sigma(\dot{c}/c)$. For this to be consistent with a rate of return of 100 percent on investments in capital, consumption must grow forever at the exponential rate $(1-\rho)/\sigma$. Because output is linear in Z, this is feasible if Z grows at the same exponential rate and consumption is proportional to Z.

To verify that this is an equilibrium, consider the problem faced by a representative consumer. At time t, the consumer will receive labor income equal to $L(\partial Y/\partial L) = GLM(t)$ and rental income on capital equal to $RZ(t)$. The consumer takes the interest rate $R=1$ as given and takes the path for labor income over time as exogenously given. The consumer chooses how much to consume and the rate of accumulation $\dot{Z}$. Since the total mass of identical consumers is 1, the aggregate rate of accumulation will also equal $\dot{Z}$. Just as in the static model, the equilibrium condition in the market with monopolistic competition is that the range of inputs produced at time t must satisfy $M(t) = Z(t)$.

Each individual consumer takes the path for $M(t)$ as given because it depends on the aggregate savings decisions of all consumers in the economy. In this sense, $M(t)$ behaves just like a positive externality, like a form of anti-smoke. Using the approach described in my 1986a paper for calculating dynamic equilibrium problems with a path like $M(t)$, which atomistic agents take as given but which is endogenously determined, it is easy to verify that the solution to the consumers problem is indeed to choose $c(t)$ and $Z(t)$ so that they grow at the rate $(1-\rho)/\sigma$. (For example, in the logarithmic case $\sigma = 1$, the equilibrium value of $c(t)$ is $c(t) = (G + \rho)Z(t)$. Substituting this and the expression $Y(t) = Z(t)(G+1)$ into equation (12) shows that c and Z grow at the rate $1-\rho$.)

One can verify directly that this equilibrium is suboptimal. Relative to the maximization problem faced by each consumer, a social planner would not take the path of wages or $M(t)$ as given; instead, the planner would take account of the fact that a higher rate of savings leads not only to higher investment income but also higher labor income. The planner would also produce more output for given Z by setting $\bar{x}$ and M at the (first-best) optimal levels rather than at the equilibrium levels. Both these effects cause the first best optimum to have a higher rate of investment and a higher rate of growth. All individuals in this economy could be made better off by a binding agreement to invest and save more than is privately optimal and to subsidize the production of a wider range of goods.

In my related paper (1986b), I argue that it is not an accident that the analysis of this equilibrium so strongly resembles one with a positive externality. This apparent "external economy" associated with the specialization is closely related to the intuition behind Marshall's use of the term. This model is not one with a true positive externality, but it nonetheless behaves exactly as if one were present.

The analysis of the dynamic equilibrium with the same preferences and cost function h, but with $g(x) = x^{\alpha}$ is quite similar. The only important difference is that the equilibrium value of R, while still constant, differs from the previous value of 1. Consumption and the stock of Z will still grow at a constant rate (though one that is algebraically more complicated to express.) The equilibrium is still suboptimal, growing more slowly than the first best optimum. Even though the static equilibrium is efficient for given a level of Z, the dynamic equilibrium offers individual agents a return from savings that is too small, and Z grows too slowly. The only intervention needed to achieve the optimum in this special case is a subsidy to savings.

In both of these equilibria, the economy will behave as if there is a form of exogenous, labor augmenting technological change. In the second case this is easy to compare with standard Cobb-Douglas descriptions of growth. Equations (7) and (8) imply that both $N(t)$ and $M(t)$ are proportional to $Z(t)$. Using output written in terms of L, M, and N as in equation (4), and impounding all the constants into a new constant A, output at time t can be written as

$$(13)\quad Y(t) = M(t)^{1-\alpha}(L^{1-\alpha}N(t)^{\alpha})$$
$$= AZ(t)L^{1-\alpha}.$$

In equilibrium, labor's share in total income is $1-\alpha$ and capital's share is α, despite the fact that the true coefficient on Z is 1. A 1 percent increase in the stock of Z causes a 1 percent increase in income, a fraction α of which is returned as payments to capital. The remaining $1-\alpha$ percent increase shows up as increased wages for labor, so labor receives the surplus arising from the apparent increasing returns. Since the rate of return on capital does not decrease with the level of the capital stock, growth can continue indefinitely. Each individual agent takes the path for $M(t)$ as given, so viewed from the aggregate level, the evolution of this economy will appear to be governed by a Cobb-Douglas technology and exogenous technological change. But any change that leads to an increase in savings—for example a tax subsidy, a decrease in the rate of impatience ρ, or a decrease in the intertemporal substitution parameter σ—will cause growth to speed up; the rate of exogenous technological change will appear to increase.

REFERENCES

Arrow, Kenneth J, "The Economic Implications of Learning by Doing," *Review of Economic Studies*, June 1962, *39*, 155–73.

Dixit, Avinash K., and Stiglitz, Joseph, "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Ethier, Wilfred J., "National and International Returns to Scale in the Modern Theory of International Trade," *American Economic Review*, June 1982, *72*, 389–405.

Hart, Oliver D., "Monopolistic Competition in the Spirit of Chamberlin: A General Model," *Review of Economic Studies*, October 1985, *52*, 529–46.

Judd, Kenneth L., "On the Performance of Patents," *Econometrica*, May 1985, *53*, 567–85.

Lucas, Robert E., Jr., "On the Mechanics of Economic Development," prepared for the Marshall Lecture, May 1985.

Marshall, Alfred, *Principles of Economics*, London: MacMillan Press, 1977.

Perloff, J. and Salop, S., "Equilibrium with Product Differentiation," *Review of Economics Studies*, January 1985, *52*, 107–20.

Prescott, Edward C. and Boyd, John, "Dynamic Coalitions, Growth, and the Firm," in E. C. Prescott and N. Wallace eds., *Contractual Arrangements for Intertemporal Trade*, Vol. 1, Minnesota Studies in Macroeconomics and Time Series Analysis, Minneapolis: University of Minnesota Press, forthcoming 1987.

Romer, Paul M., "Dynamic Competitive Equilibria with Externalities, Increasing Returns, and Unbounded Growth," unpublished doctoral dissertation, University of Chicago, 1983.

______, (1986a) "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–38.

______, (1986b) "Increasing Returns, Specialization, and External Economies: Growth as Described by Allyn Young," Working Paper No. 64, Rochester Center for Economic Research, December 1986.

Sattinger, M., "Value of an Additional Firm in Monopolistic Competition," *Review of Economic Studies*, April 1984, *51*, 321–32.

Schmitz, James, "Optimal Growth and Product Innovation," Department of Economics Working Paper, University of Wisconsin, October 1986.

Stokey, Nancy L., "Learning-by-Doing and the Introduction of New Goods," Working Paper No. 699, Center for Mathematical Studies in Economics and Management Science, Northwestern University, September 1986.

Vassilakis, Spyros, "Increasing Returns and Strategic Behavior," unpublished doctoral dissertation, Johns Hopkins University, 1986.

Dynamic Coalitions: Engines of Growth

By EDWARD C. PRESCOTT AND JOHN H. BOYD*

In this study we consider an equilibrium model with sustained growth, in which a dynamic coalition production technology plays a key role. The technology has three major implications. First, even without exogenous technological change, there is sustained growth in per capita output and consumption. Second, unlike the neoclassical growth model, policy can have important effects on *average* growth rates. Specifically, any policy which distorts investment-consumption decisions will alter the equilibrium growth rate. In the economy studied here, such policies are not necessary for efficiency, but in slightly different environments they could be. Third and finally, equilibrium in this model has an interesting industrial organization implication. That is, firm (coalition) size may vary cross sectionally, but there is no tendency for the size distribution to collapse on a single point or, on the other hand, to spread over time. Neither is there any tendency toward a single monopoly firm.

We believe these industrial organization implications are consistent, at least in a highly stylized way, with what is actually observed. This we think is important, for most previous studies have had difficulty in simultaneously accounting for growth and firm-size observations. The problem is, to have growing per capita output, returns to capital cannot diminish. But for the usual production technologies with a labor input, when returns to capital are constant (or increasing), there are also increasing returns to scale. Increasing returns to scale leads to a single monopoly firm, or at least precludes the existence of a competitive equilibrium.

In the environment studied here, the emphasis is on technological knowledge, which is embodied in workers and is partly organization specific. Hereafter, we refer to this knowledge as "coalition capital" for short. Physical capital is not included in the analysis, although it could harmlessly be added. Returns to investment in coalition capital are "constant" in a sense which will be made precise in the following section. This results in the possibility of long-run sustained growth without exogenous technological change.

A key assumption here is that workers' productivity depends not only on their own human capital but also on that of their coworkers. Thus, from the perspective of individual agents there is an externality: their personal human capital acquisition decisions affect the productivity of others. And, if such decisions were made in a decentralized market, the expected result would occur: namely, the rate of coalition capital formation would be too low. We do *not* have in mind, however, that this technological interdependency exists between all workers—so that when one invests in his own human capital, it shifts out the production frontier for the nation or world. Rather, such interdependencies are assumed to exist only for workers who are members of the same coalition and have previously worked together. The organizational structure of coalitions allows for richer contracting arrangements than those observed in decentralized markets, and this permits the coalition to correctly reward each member's capital investment. In fact, the above-mentioned "externality" effectively disappears because it is internalized inside the coalition. Importantly, coalitions have no monopoly power, and none can earn rents in equilibrium—they just earn the market returns on their coalition capital.

Our results are different from those obtained elsewhere in one extremely important respect. In the neoclassical model, economic advancement is determined by exogenous technological change. The model is incapable of delivering any important insights into

*Federal Reserve Bank of Minneapolis, Minneapolis, MN 55480, and University of Minnesota, Minneapolis, MN 55455. The views expressed here are our own.

how growth rates are determined or what growth rate is best. In a recent study (Paul Romer, 1986), equilibrium growth is endogenous to the model, but that growth is strictly due to external effects which are not reflected in market prices or private contracts.[1] Thus, the *laissez-faire* equilibrium growth rate may be positive, but it will generally be too low. Obviously, in these environments policy interventions will be welfare improving, at least if the government knows how to set the right policies. In our environment, by contrast, equilibrium growth is due to production externalities which are fully accounted for in coalition contracting. Thus, equilibrium growth is not only endogenous to the model, it is endogenous to the coalitions themselves. Although policy interventions can affect equilibrium growth rates, they are not necessary for efficiency.

We recognize that ours is an extreme characterization of the world and that the truth probably lies somewhere between the two extremes. It seems likely that the economic returns to investing in knowledge are partly but not entirely captured by the individual agents or groups of agents who do the investing. Yet, if such investments are genuinely important in determining equilibrium growth rates, as is widely believed, so too are the institutions and arrangements which determine growth under *laissez-faire*. The fact that there are persistent differences in growth rates across countries suggests the need for models which focus on the institutional arrangements within countries as well as those factors affecting technology. We view our exercise as a first step in that direction.

I. The Economy

Initially, there is some given number of old agents. Each period, that number of young agents are born, and they live for two periods. Thus, at all points in time there are equal numbers of young and old. Those born in period t for $t = 1, 2, \dots$ have utility function u: $R_+ \times R_+ \to R_+$:

$$u(y_t, z_{t+1}) = \ln y_t + \beta \ln z_{t+1}$$

where y_t is consumption when young, z_{t+1} is consumption when old, and β is a parameter, $0 < \beta < 1$. The utility function of an initial old agent is simply $\ln z_0$.

All production activities are carried out by coalitions of agents which are the "firms" in this economy. We have described and discussed such coalitions in some detail elsewhere (see our forthcoming paper), arguing that for some purposes they may be a better representation of the firm than is the standard one (a technology specified as a subset of the commodity space). We discuss these coalitions only briefly here.

Coalitions are groups of agents, all of whom have access to the same blueprint technology. All agents are identically endowed, but may choose to accumulate knowledge at different rates. Thus, over time coalitions may differ too, depending on the decisions of their members which jointly and cumulatively determine coalition capital. Each coalition is composed of young members and old members, all of whom employ their labor services in producing the consumption good. Also produced is the human capital of young coalition members. In the next period they will become old members, and the more human capital they carry with them, the greater the production possibilities of the coalition—both for producing the consumption good and for producing more productive workers in the future. Although individual agents only live two periods, coalitions endure forever in this economy. Formally, the technology is as follows.

Technology

A coalition is characterized by its size in terms of number of experienced old workers M and the expertise of each of its members k. The coalition has young worker inputs N and produces the consumption good as well as new capital or expertise k' which is

[1] See Romer for a review of the literature on competitive growth models with externalities. Robert Lucas (1985) also has a competitive model with sustained growth and no exogenous technological change. The key features of his model are that all capital including human capital is reproducible and that production technologies display constant returns to scale.

embodied in each of its young workers. Letting $n=N/M$ be the number of young workers per old, output of the consumption good produced per old, c, is constrained as follows:

$$c \leq kf(n)-h(n)k'n \tag{1}$$

where f is an increasing, differentiable, strictly concave, positive real-valued function and where h is an increasing, differentiable, strictly convex, positive real-valued function. Additional properties will be imposed on functions f and h that guarantee the existence of an equilibrium with positive growth.

It is important to note that the output of the consumption good c, like n and k, is per experienced member, while k' is expertise per young worker. In equilibrium all coalitions will select the same c, n, k' if they have the same k, *independent* of their coalition size. Consequently there is no "optimal" coalition or firm size to which coalitions converge. Rather, in equilibrium size differences persist with no tendency for big coalitions to become smaller or small coalitions to become bigger.

The rationale for this particular production constraint is as follows: If investment were zero (i.e., $k'=0$), the output of the consumption good would be $kf(n)$, with $f(\cdot)$ being a standard strictly convex production function and k the "technology parameter." It is important that this function not be homogeneous of degree less than one in k, for then there cannot be sustained growth. As will be seen, the problems of "increasing returns" associated with $kf(n)$ are finessed by the dynamic coalition mechanism.

The second part of the constraint is the investment in new expertise. This output $k'n$ is costly in terms of output of the consumption good. The key assumption that differentiates this model from our earlier one (forthcoming) is that here, as the number of young per old increases, the cost of a given investment increases—that is, $h(\cdot)$ is an increasing function. This we think is a reasonable assumption, for it implies with more young workers per experienced worker, expertise transfer and enhancement become increasingly costly in terms of the current consumption good.

Absent borrowing and lending between coalitions consumption at a given date is constrained as follows:

$$z+ny \leq c \tag{2}$$

with y being consumption of young and z consumption of old. We will show that in equilibrium there will be no borrowing and lending between coalitions and consequently that borrowing and lending markets need not be included.

II. Constant Growth Equilibrium

We seek a constant growth equilibrium. In this context, *constant growth* means that the capital stock and the consumption of young and old all grow at a common (gross) rate x. Unlike the neoclassical growth model's balanced or steady-state growth path, which is independent of initial capital, our steady-state growth path is proportional to the initial coalition capital k_0. Summarizing the desired properties of constant growth:

$$k_t = k_0(x^*)^t \tag{3}$$

$$y_t = y^* k_0(x^*)^t \tag{4}$$

$$z_t = z^* k_0(x^*)^t. \tag{5}$$

Equilibrium elements (x^*, y^*, z^*) are to be determined.

The key equilibrium condition is that the consumption-training mix offered young (i.e., the (y, k') pair) must be competitive in terms of the lifetime utility that the young will realize. It is competition for young workers by existing coalitions that determines the equilibrium allocation. The old maximize their consumption

$$z_t = \max_{n_t, k_{t+1}, y_t} \{k_t f(n_t) - h(n_t)k_{t+1}n_t - y_t n_t\} \tag{6}$$

subject to the (y_t, k_{t+1}) yielding at least the market indirect utility value for the young members attracted. This constraint is

$$\text{(7)}\quad \ln y_t + \beta \ln(z^* k_{t+1}) \geq \ln(y^* k_t) + \beta \ln(z^* k_t x^*) = u_t^*$$

or

$$\text{(8)}\quad \ln \frac{y_t}{k_t} + \beta \ln \frac{k_{t+1}}{k_t} \geq \ln y^* + \beta \ln x^*.$$

A final equilibrium condition is that the labor market clear. As there are equal numbers of young and old workers, this requires that

$$\text{(9)}\qquad n_t^* = 1$$

for all t.

Letting $x = k_{t+1}/k_t$ and $y = y_t/k_t$, the optimization problem is

$$\text{(10)}\quad z^* = \max_{n,x,y \geq 0} \{ f(n) - h(n)nx - yn \}$$

subject to

$$\text{(11)}\quad \ln y + \beta \ln x \geq \ln y^* + \beta \ln x^*.$$

The first-order conditions for this program when evaluated at equilibrium values $x = x^*$, $y = y^*$, $n = n^* = 1$ are

$$\text{(12)}\qquad 1 = \lambda / y^*$$

$$\text{(13)}\qquad h(1) = \lambda \beta / x^*$$

$$\text{(14)}\quad f'(1) - h'(1)x^* - h(1)x^* - y^* = 0$$

where λ is the Lagrange multiplier. Solving these necessary first-order conditions of this (nonconcave) program yields

$$\text{(15)}\quad x^* = \frac{\beta f'(1)}{\beta h'(1) + \beta h(1) + h(1)}$$

$$\text{(16)}\quad y^* = \frac{f'(1)h(1)}{\beta h'(1) + \beta h(1) + h(1)}.$$

Substituting these values along with $n^* = 1$ in (10), we obtain

$$\text{(17)}\quad z^* = [f(1)[\beta h'(1) + \beta h(1) + h(1)] - f'(1)h(1)(1+\beta)] / [\beta h'(1) + \beta h(1) + h(1)].$$

These elements are all nonnegative given our assumptions. By multiplying function h by a positive constant, y^* and z^* are unchanged, but x^* is multiplied by the reciprocal of that constant. Thus, for a suitable h, x^* will exceed one and there will be positive growth. We assume that h is such that this is the case.

Some additional conditions are needed to ensure that in equilibrium it is not in the interest of coalitions to borrow from and lend to each other. In particular, we want it *not* to be in the interest of a borrowing coalition to make a larger per capita investment in coalition capital and a lending coalition to make a smaller one. If this were to happen, coalitions would not remain identical in equilibrium and the above (x^*, y^*, z^*) would not define a constant growth competitive equilibrium.

A condition which assures that there are no gains from concentrating the capital in a fraction of the population is as follows:

$$\text{(18)}\quad \max_{n,x \geq 0} \{ f(n) - h(n)nx + q^*nx - (q^*x^*/\beta) - w^*n \} \leq 0$$

where

$$\text{(19)}\qquad w^* = y^* + q^*x^*$$

$$\text{(20)}\qquad q^* = h(1).$$

In the above, w^* is the equilibrium real wage divided by k and q^* the equilibrium price of new capital for an economy in which capital is tradeable. This technical issue is developed fully in Prescott (1986).

III. Concluding Remarks

This model, like those of Robert Lucas (1985) and Romer accounts for sustained growth in per capita income with little if any tendency for countries to converge to a common growth path. The hope, however, is that this structure will prove useful in accounting not only for similarities, but also for differences in growth experiences, some of which are dramatic. Our theory predicts more rapid growth rates if the fraction of resources allocated to enhancing coalition capital is larger. This model also implies that young residents of low-income countries will gain income by moving to a high-income country. Perhaps improved time-allocation studies of people at work in organizations will confirm or refute the value of this abstraction.

REFERENCES

Lucas, Robert E., Jr., "On the Mechanics of Economic Development," prepared for the Marshall Lecture, Cambridge University, May 1985.

Prescott, Edward C., "Dynamic Coalitions, Growth and the Firm: An Addendum," Working Paper, Minneapolis Federal Reserve Bank, May 1986.

______ **and Boyd, John H.,** "Dynamic Coalitions, Growth, and the Firm," in E. C. Prescott and N. Wallace, eds., *Contractual Arrangements for Intertemporal Trade*, Volume 1, Minnesota Studies in Macroeconomics and Time Series Analysis, Minneapolis: University of Minnesota Press, forthcoming.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*, 1002–38.

THE FAMILY[†]

What Have We Learned from the Economics of the Family?

By ROBERT J. WILLIS*

The family is distinguished from other social institutions, such as firms, by its crucial role in the production and nurture of children and its rationale is ultimately to be found in the preferences of individuals for own children. Sexual reproduction means that the production of one's own child requires the participation of another person of the opposite sex. The production of a child who will survive, become a successful adult, and produce his or her own children requires the expenditure of both personal and purchased resources over a lengthy period of time.

Although interesting insights on the family can be culled from the classics, systematic development of the economics of the family is a recent phenomenon, beginning in the late 1950's when Harvey Leibenstein (1957) and Gary Becker (1960) attempted to address the determinants of fertility behavior within the framework of consumer theory. In this paper, I provide a brief overview of the history of family economics since 1960 and, along the way, offer a selective assessment of what has been learned from it.

I attempt this assessment by asking how far we have progressed in answering a few of the larger theoretical, empirical, and policy questions that have motivated economists' interests in an area customarily studied by sociologists and demographers, or that have caused economists dealing with more traditional subject matter to incorporate the family into their work. Among the set of questions that have been addressed within the literature during the past twenty-five years are:

1) What are the causes of the historical association between economic growth and development and demographic transition from high to low levels of fertility and mortality? Of what relevance is the historical experience of currently developed countries to contemporary LDCs? Should fertility reduction be a primary goal of policy in the developing countries? Are the developed countries in danger of extinction because of fertility below replacement levels?

2) What was the cause of the post-World War II "baby boom" and subsequent "baby bust"? Was the baby boom a one-time aberration from a secular decline in fertility, or can we expect substantial fluctuations in the birth rate in the future? What are the consequences of the baby boom for the economic welfare of cohorts born during and after the boom?

3) Is the traditional family "dead" in the United States and other developed countries? Why did the divorce rate double in a decade? Why the growth in female-headed households? Why do so many divorced fathers fail to support their children? Has the sexual division of labor within the family changed as a consequence of the growth of female labor supply? To what extent are these changes in the family caused by social policy, and to what extent are they a product of basic market forces associated with modern economic development? What are the consequences of these changes for the welfare of future generations?

I attempt to touch on some issues from each of the three areas in which the questions are grouped. However, constraints imposed by limitations of space, time, and most im-

[†] *Discussants*: Kenneth Wolpin, Ohio State University; Robert Pollak, University of Pennsylvania; T. Paul Schultz, Yale University.

*University of Chicago, and Economics Research Center/NORC, 1155 East 60th Street, Chicago, Il 60637–2799.

portantly by limitations of my knowledge prevent me from fully addressing in this paper all of these questions or others that might well have been added.

I. The 1960's

In his first paper on fertility determinants, Becker (1960) suggested that parental demand for children could be treated as analogous to the demand for producer or consumer durables, depending on whether parents expected net pecuniary returns from children or received direct utility from them. The most important analytical contribution of this paper was Becker's hypothesis that the cost of children was in part endogenous because parents receive utility from increased child "quality" as well as from increased numbers of children. This hypothesis provided a partial rationale for the empirical observation that family size often tended to be negatively related to family income without recourse to the assumption that children are inferior goods. Thus, an increase in total expenditures on children caused by increased family income might be devoted largely to increased expenditure per child rather than to increases in the number of children, just as an increase in income might lead a consumer to shift from an economy car to a luxury car rather than increase the number of cars.

The development of family economics during the 1960's was closely linked to several other rapidly growing areas of economics. Most notable among these were life cycle theories of consumption (Franco Modigliani and Richard Brumberg, 1954) and human capital (Becker, 1964) and static theories of labor supply (H. Gregg Lewis, 1957; Jacob Mincer, 1962), household production and time allocation (Becker, 1965) and the characteristics approach to consumption (Kevin Lancaster, 1966).

II. The Schultz Volume

The economics of the family emerged as a distinct subfield with the publication in 1973 and 1974 of two special issues of the *Journal of Political Economy* which were reprinted in Theodore Schultz (1974).[1] Papers in this volume consolidated the theoretical work of the previous decade, struck off in new theoretical directions, and began to address the empirical content of the theory with the aid of large-scale micro data sets and new econometric methods.

In the first category, I (1974) presented a model of fertility behavior which synthesized the quality-quantity model suggested by Becker (1960) with a model of household production and human capital investment emphasizing the role of female time allocation between market and home work based on the earlier work of Becker (1964, 1965) and Mincer (1963). This model provides two possible reasons for the negative relationship between income and fertility which may be labelled, respectively, the "female cost of time hypothesis" and the "quality-quantity interaction" hypothesis.

The cost of time hypothesis follows from the assumption that childrearing is relatively more intensive in the use of mother's time than are non-child-related household production activities. When a wife does not engage in market work, I (1974) showed that the shadow value of her time, and hence the marginal cost of children, is an increasing function of husband's income and when women do participate in the market, the cost of time is determined by her (marginal) wage rate. Since the wife's time allocation is endogenous, the model also provides an explanation for the negative correlation between the presence of young children and female labor supply. The cost of time hypothesis suggests that the fertility demand function should distinguish between male and female wages and, in reduced form, that there should be interactions between male and female wages. I (1974) and W. G. Sanderson and I (1971) provided empirical evidence for the interaction effect using U.S. census data.

[1] For convenience, citations to these papers will be to the Schultz volume rather than to the specific *JPE* issue.

Both I (1974) and Becker-Lewis (1974) showed that an important implication of the quality-quantity model had been overlooked in Becker's 1960 paper. Specifically, because quality and quantity enter multiplicatively into the household budget constraint, they show that variation in the household's choices of the number and quality of children caused by changes in income or the cost of mother's time induces endogenous changes in the marginal cost of the number and quality of children. For example, if the income elasticity of demand for quality exceeds that for quantity, an income increase will tend to increase the marginal cost of quality relative to quantity, thereby inducing a substitution effect against number of children which may partially or more than offset a positive income effect in favor of children. As I shall explain later, recent work on altruistic preferences and intergenerational transfers has lead to some interesting reinterpretations of the quality-quantity model.

The scope of the economic theory of the family was significantly expanded in two other papers in the Schultz 1974 volume. First, Becker introduced a theory of marriage which considers the sources of gains to marriage and formalizes the concept of a marriage market. In his theory, the gains to marriage arise because of gains to the division of labor in household production, because of the joint production of "own children" and other marriage-specific capital and, finally, because of altruistic utility interactions between the marital partners which represent the role of "love." Given this microfoundation, Becker builds a theory of a competitive marriage market in which individual males and females are matched. He shows that if the joint marital income can be costlessly redistributed between the partners, there will exist a Pareto optimal competitive equilibrium which maximizes the average gain to marriage across the entire population. He also examines patterns of assortative mating in the equilibrium, showing that, under simplifying assumptions, "likes marry likes" when husband's and wife's traits are complementary and "unlikes" marry when they are substitutes.

The second new theoretical direction in the Schultz 1974 volume was provided by Marc Nerlove who suggested that the economic theory of the family could be used as the micro-foundation of a "new Malthusian" theory of population and economic growth. While Nerlove's paper only outlined some of the elements of the theory, it foreshadowed the development of a substantial body of theoretical work during the 1980's to which he among others have made substantial contributions that I will discuss later.

Finally, a number of papers in the Schultz 1974 volume began to address the empirical content of the theory of the family with the aid of large scale micro data sets and new econometric methods. Especially noteworthy were papers by Reuben Gronau and James Heckman which broke new ground in econometric methodology by addressing the question of sample censoring in the estimation of the value of time for nonworking women and in recovering household preference parameters from market data. Again, this work foreshadowed the later development of important new econometric methods by Heckman and others to deal with censoring, self-selection, and longitudinal data issues that arise in models of family behavior. One of the most interesting substantive findings in the volume was by Mincer and Solomon Polachek, who found that interrupted work careers for married women may account for a substantial fraction of the male-female wage gap.

III. Two Schools of Family Economics and the Baby Boom

The theoretical approach represented in the Schultz volume was labelled the "new home economics" by an insider (Nerlove) and the "Chicago School" theory by a variety of outsiders, both within economics and from other disciplines such as demography and sociology in which family and population studies traditionally reside. At the time, the main rival to the Chicago School theory within economics was represented by Richard Easterlin (1969), who was developing an alternative theory which attempted to synthesize economic and sociological approaches to fertility.

Most notably, Easterlin (1973) sought to explain the postwar baby boom and subsequent baby bust by shifts in preferences for children caused by changes in intergenerational relative income across cohorts. He argued that the desired standard of living of young adults is shaped by the living standards they experience while growing up. If current income is high (low) relative to this standard, they will tend to marry early (late) and have high (low) fertility. According to this hypothesis, cohorts born during the Great Depression and World War II who entered their childbearing years during the postwar boom felt that they could afford large families, while their children, reared in more affluent times and faced by severe labor market competition caused by their large numbers, would tend to delay fertility. To the extent that relative cohort size negatively influences cohort incomes, his hypothesis implies a continuation of endogenous fertility swings into the future.[2]

Some in the Chicago School viewed any argument based on shifting preferences with suspicion. For example, George Stigler and Becker (1977) argued on methodological grounds that stable preferences should always be posited, although they would permit variations in household behavior to be "explained" with variations in household production technology as well as variation in prices and income. While acknowledging the heuristic value of the distinction between tastes and technology in the household production model, two members of the "Pennsylvania School," Robert Pollak and Michael Wachter (1975), argued (perhaps too strongly) that in practice it is impossible to disentangle variation in behavior caused by these two sources with data on observed household behavior.

The challenge of explaining the postwar baby boom and subsequent baby bust became, for a time, the primary testing ground between the rival approaches. Several studies using aggregate econometric models of postwar fertility behavior have attempted to test both the Easterlin and the Chicago school models with mixed results. For example, R. D. Lee (1977) found support for the relative income hypothesis while William Butz and Michael Ward (1979) obtained results favoring the Chicago model. Indeed, they went so far as to claim that, because of growth in female labor force participation, the female wage effect may have become sufficiently strong to cause "countercyclical" fertility.

Considerable skepticism is warranted about the possibility that aggregate data from one cycle is sufficient to identify either model. Moreover, there is little quantitative agreement between cross-section and time-series estimates of either model. For example, I (1974) completely failed to account for changes in cohort fertility on the basis of my cross-section estimates whereas Butz and Ward's time-series estimates of the Willis model appears to be quite successful in fitting the fertility swing. By the same token, Easterlin (1973) and Lee find that fertility is quite sensitive to variations in relative income measures in aggregate data while Yoram Ben Porath (1975) finds almost no effect of relative income using micro data.[3]

The division of the economics of the family into rival schools was short-lived. As Sanderson (1976) argued, the two schools had much in common. For instance, both emphasize the importance of household resource constraints in explaining demographic behavior. Additionally, it is interesting to note that, despite the position he takes in Stigler and Becker, Becker's own work increasingly stresses theorizing about preferences including intergenerational linkages within the family (for example, Becker and Robert Barro, 1985). Finally, as the field developed, it attracted new researchers, problems, and approaches not associated with either school.

[2] More recently, Finis Welch (1979), M. C. Berger (1985), and others have found evidence of significant cohort-size effects on earnings.

[3] For a more detailed methodological discussion of this work, see T. P. Schultz (1981).

IV. Biology and Fertility Dynamics

As an example of commonality, both Easterlin (1978; Easterlin, Pollak and Wachter, 1980) and members of the Chicago School (Robert Michael, 1974; Michael and I, 1975; Heckman and I, 1976), argued that economists must face the "facts of life" by incorporating the biological aspects of reproduction into models of fertility. That is, to use Easterlin's terminology, it is necessary to analyze the "supply" as well as the "demand" for births. This lead economists to draw on important developments in mathematical demography and biostatistics in which reproduction is represented as a Markov renewal process (for example, E. Perrin and M. C. Sheps, 1964), and suggested the need for a shift from a static to a dynamic framework in economic theories of fertility.

The biology of reproduction implies that fertility decisions are inherently sequential and stochastic. Couples cannot directly choose to have or not have a birth at a given time. Rather, fertility is controlled indirectly by actions such as coital frequency, contraception, and breastfeeding, which determine the risk of pregnancy, and by other actions, such as abortion, which determine whether a pregnancy will result in a live birth. Recognition of these features of the reproductive process led Heckman and me to an attempt to formulate a rational model of reproductive decision making within a stochastic dynamic programming framework, and to recast the empirical framework in which fertility models are tested with data on birth intervals instead of children ever born. While the logic of this dynamic approach to fertility behavior is appealing, it presents serious technical problems in obtaining closed-form solutions to the dynamic program, unambiguous theoretical predictions, or estimable structural econometric models.

Research on dynamic fertility models has been and continues to be an active area of research. Although several attempts have been made (with limited success) to generate more operational dynamic theoretical models (for example, V. J. Hotz and R. Miller, 1985; J. Newman, 1984), most of the activity has been devoted to the development of econometric methods (for example, Heckman and B. Singer, 1984) and to estimates of reduced-form models of birth spacing. In the latter category, several papers indicate that imperfect fertility control and changing opportunity costs over the life cycle play a significant role in explaining fertility behavior (Mark Rosenzweig and Schultz, 1985; Hotz-Miller; Heckman and J. R. Walker, 1986). In addition, P. A. David and T. A. Mroz (1986) use such a model to provide evidence of deliberately controlled fertility before the French Revolution. Another promising line of research by Kenneth Wolpin (1984) attempts to combine theory and structural estimation by formulating a dynamic programming model which is solved numerically in each iteration in the maximization of a likelihood function.

V. Family Instability: Out-of-Wedlock Births, Divorce, and Child Support

To many Americans, the growing instability of the family has been one of the most troubling social developments during the past twenty years. The symptoms of breakdown abound. While fertility in general has fallen, the rate of out-of-wedlock childbearing by teenagers has grown. The divorce rate doubled within the decade 1965–75 and has remained at high levels. Growth in female-headed households has been implicated as a major source of poverty, in part because many fathers fail to pay child support. The economics of the family provides some insights into the underlying causes of family instability, although much work remains to be done on both theoretical and empirical fronts. This section briefly outlines a few of the more interesting contributions in this area from the family economics literature.

A natural starting point for considering out-of-wedlock childbearing, divorce, and nonsupport of children is to ask why long-lasting marital unions have been the norm in so many times and places. Part of the answer stems from the desire of individuals for own children. Suppose, following Yoram Weiss and myself (1985), that an individual's utility depends on his or her own consumption and on the number and welfare of own children. Since a male and female must cooperate to produce a child, the child's welfare is a collective good from the viewpoint of the parents.

The collective goods nature of children provides a strong rationale for the traditional strategy of first marrying and then having children. Although it is feasible for a woman to have a child without the knowledge of the father, typically it would not be in her interest to do so because, since both parents benefit from the child's welfare, a Pareto optimal allocation of their joint resources requires both to contribute to the child's welfare. Moreover, because of the potential of free riding, an optimal resource allocation is most easily attained when the partners can monitor one another. If monitoring is not easy, as when one partner has custody, the Weiss-Willis model shows that the shadow price of the child's welfare increases to both partners resulting in underexpenditure on the child. Finally, to the extent that marriage entails enforceable contractual elements, a women is clearly in a better bargaining position if she delays childbearing until after marriage.

The most obvious hypothesis to account for the growth of out-of-wedlock childbearing is the growth of AFDC welfare programs in which the state replaces the father as a source of child support. Despite the strength of the theoretical case for this hypothesis, very little empirical support for it has been found (for example, M. J. Bane and D. Ellwood, 1984). Two recent studies suggest, however, that AFDC effects may be found when empirical analysis is more closely guided by theory than is the case in many of the earlier studies. In one, A. Leibowitz and M. Eisen (1986) argue on theoretical grounds and find empirical support for the hypothesis that AFDC need not affect the rate of teen pregnancy when abortion is available, but that higher AFDC benefits increase the likelihood that a pregnant teen will decide to keep her child. In the other, M. S. Bernstam and P. L. Swan (1986) emphasize the theoretical importance of the competing roles of father's and state resources in determining childbearing by teenagers. They find significant AFDC effects using state data which controls for the earnings potential of young males.

Most Americans do marry and most children are produced within marriage. However, the rapid growth of divorce means that many of these children will spend at least part of their childhood within a female-headed household and will be exposed to a substantially higher risk of poverty during that period (Bane and Ellwood, 1983; G. J. Duncan and S. D. Hoffman, 1985).

The economic theory of divorce is the obverse side of the theory of marriage. Becker, Landes, and Michael (1977) emphasize the role of uncertainty and imperfect information in accounting for divorce. Thus, while expected gains to the partnership are assumed to be positive at the time of marriage, there is some probability that the net gains will turn out to be negative *ex post* either because the partners are not well suited to one another, or because superior opportunities outside the marriage arise for one or the other partner. Applying the Coase Theorem, they argue that divorce will not occur when one partner would be better off breaking the marriage while the other partner suffers if the latter can compensate the former by a reallocation of income within marriage.

This proposition has received empirical support in a study of the effects of the shift from fault to no-fault divorce laws by Elizabeth Peters (1986). She finds no effect of the type of law on the probability of divorce, but does find that divorce settlements received by women are smaller in no-fault states. The latter finding is consistent with a hypothesis developed in Weiss-Willis which predicts that implicit *ex ante* marriage contracts will provide "divorce insurance" for the custodial parent (almost always the mother) as part of the divorce settlement; no fault removes an important *ex post* source of enforcement of such a contract.

Apart from the suspicion that growth in income transfer programs play a role in contributing to divorce, especially among low-income groups, the main line of theoretical explanation for the increase in the divorce rate focuses on basic economic forces which cause a decrease in the gains to family life and lead to a shift in the sexual division of labor within the family.

In household production models (for example, Becker, 1965), the family is conceived, in part, as a productive organization in which nonmarket commodities are produced with purchased goods and the time of household members. Efficiency dictates a division of labor within the household and between household and market activities according to the comparative advantage of its members. In an important addition to this argument, Becker (1981; 1985) demonstrates that, because of increasing returns to the rate of utilization, initial differences in comparative advantage will tend to be magnified by optimal investments in skill-specific human capital. Thus, he argues, intrinsic sex differences in the biology of reproduction (or discrimination against women) tend to induce reenforcing investments in the productive skills of men and women, respectively, in market and nonmarket production.

The theory helps to explain the sharp sexual division of labor within the "traditional family" and points to reasons why the division of labor may become less marked as economic development causes an increase in the comparative advantage of market relative to nonmarket modes of production and as market discrimination against women decreases. Decreased specialization in household production together with reduced demand for children, both caused by the rising market value of female time, tend to contribute to a reduction in the gains to marriage and hence to an increase in the probability of divorce. Moreover, increased divorce risk may tend to feed on itself by increasing the pool of eligibles for remarriage and by leading women into precautionary investments in marketable skills as self-insurance against divorce.

Despite the pressures for changes in the sexual division of labor, many observers (even including Becker, 1985) have been struck by the persistence of traditional sex roles within the family in societies and by the failure of intrafamily time allocation to adjust to the growth of female labor force participation in societies as diverse as the United States, Japan, and the Soviet Union. Microeconomic studies do document some responsiveness in time allocation between husbands and wives in response to variations in husband's and wife's wages (for example, Gronau, 1977), but the magnitude of the variation appears to be modest. The failure of men to take over more responsibility in the home is sometimes suggested as a reason that career-minded women are delaying marriage and childbearing.

To date, few empirical studies have attempted to test such explanations of the divorce explosion. In one, Michael (1986) finds evidence of effects on divorce of male and female wages and of an index of contraceptive technology in a relatively unstructured time-series analysis, but he cautions that the evidence is based on only a single episode. More generally, it has proven difficult to find enough well-measured exogenous variables to permit cause and effect relationships to be extracted from correlations among factors such as the delay of marriage, decline of childbearing, growth of divorce, and increased female labor force participation with aggregate or even micro level data.

VI. Investment in Children, Children's Welfare, and Parental Altruism

In every culture, the family is the primary agent responsible for the care and nurture of children and "family background" is typically found to be one of the most powerful predictors of adult achievement or failure. The conceptual framework for treating the relationship between the family and the welfare of its children has been greatly elaborated since Becker (1960) introduced the term "child quality," measured by the level of expenditure per child, as an object of parental preference. Much of the elaboration is due to Becker.

In his Woytinsky Lecture (1967), Becker argues that the distribution of income from investment in human capital reflects the interaction of variations in the demand for investment and supply of funds across individuals in the population. He emphasizes that the family may influence demand through its effects on a child's capacity to benefit from investment and influence supply, assuming the child faces borrowing constraints, by its willingness to help finance the investment.

Subsequently, the demand side of this model was recast in terms of Becker's 1965 household production framework in several papers in the Schultz 1974 volume (myself; Leibowitz; D. N. De Tray). These papers assume that child quality is produced with inputs of parental time and purchased goods according to a household technology which may vary with parental characteristics, with the child's innate traits, and with environmental factors. Empirically, Leibowitz focused on inputs, finding that more educated mothers tend to devote more time to child care and are less likely to drop out of the labor market, while De Tray used state data in an attempt to study educational attainment as an output measure.

An unresolved issue in these papers concerns the operational definition of child quality. Is it a scalar, or do the myriad of possible child characteristics such as sex, health, intelligence, personality traits, educational attainment, etc. enter the parent's utility function as separate arguments? The gain in generality of the latter approach is offset by its greater theoretical complexity and by lack of data, although models of sex preference have gone in this direction in part because gender is easily observed (Ben Porath and Welch, 1976).

An alternative strategy stems from Becker's 1974 approach to altruism and Barro's 1974 demonstration that parental preferences can be represented within the overlapping generations framework by a "dynastic utility function" if parents' utility is equal to the sum of their own utility from consumption and the lifetime utility of each of their children multiplied by a weight representing the degree of parental altruism. Because of recursivity, the parents' utility is equal to the sum of the levels of utility from own consumption of their children, grandchildren, and all subsequent generations in the dynasty discounted by the rate of altruism. Given such preferences, a child's level of lifetime utility (i.e., the weighted sum of his utility from consumption and the utility of his children) can be interpreted as a scalar measure of child quality.

Following this strategy, Becker and Nigel Tomes (1976, 1984) address the determination of the optimal investment in children. Their analysis provides an explicit model of the role of the family in the finance of human capital which had only been discussed informally in the Woytinsky Lecture. To illustrate their approach, consider the following simple example. Suppose that parents are altruistic toward their children, that they face a perfect capital market, and that they can enforce any distribution of family income among family members they desire. Also assume that children cannot borrow to finance their own consumption or human capital investment while they are young.

Taking the number of children as given, parental utility maximization can be represented as a two-stage process in which parents first allocate resources so as to maximize total family wealth by equating the rate of return to investment in human capital for each parent and each child to the rate of interest and, second, distribute the (maximized) wealth among family members so as to maximize a weighted sum of their utilities with the weights determined by the degree of altruism. The level of investment per child is fully determined by the rate of interest and marginal return schedule for human capital, which might vary across children because of innate ability differences. In order to maximize total family wealth, parents reenforce innate ability differences by investing more in children with higher return schedules. Assuming that parents' preferences are egalitarian (i.e., they give equal weight to the utilities of each child) parents will make asset transfers so as to equalize the wealth of each child and, hence, equalize their levels of lifetime utility.

The distribution of wealth between parents and children depends on the degree of parental altruism and, given the degree of altruism, the direction and magnitude of intergenerational transfers within the family depends on the relative (maximized) earnings potentials of parents and children. For example, relatively wealthy and highly altruistic parents may make positive "bequests" to all children, with more able children receiving smaller bequests. In the converse case, parents may desire asset transfers from their children in order to "repay" in part the parents' cost of childrearing and investment in children's human capital. In this case, redistribution of wealth among children requires asset transfers from more to less able siblings. Transfers from children to parents and among siblings pose obvious enforcement problems which have sparked an interest in issues of "the economics of intergenerational control" (Donald Parsons, 1982).

Optimal fertility is determined when the marginal cost of an additional child, measured by the parents' net expenditure on her or him, is equal to the monetary value of the marginal utility of a birth to the parent, measured by the child's lifetime utility divided by the parents' marginal utility of wealth. A "low-quality" child who receives little investment in human capital thus may be a "high-cost" child if he has wealthy parents who make a large bequest to him and, conversely, a highly educated child of poor parents may be of lower cost because he provides them with sizeable "old age transfers." The net expenditure on a child plays a dual role in this model. It is both an endogenously determined shadow price which induces quality-quantity interactions in the demand for children similar to those described earlier, and it is a measure of the net intergenerational transfer from the older to the younger generation within families.

The model in the example described above contains very strong assumptions about markets, about the form of parental preferences, and about the nature of intrafamily relationships, and it provides a number of strong and testable predictions at both micro and macro levels of analysis. As such, it provides a point of departure for a wide variety of alternative models of the family in which one or another of these assumptions are relaxed or altered and in which alternative empirical hypotheses can be generated about a number of different dimensions of behavior. Space constrains me to mention only a few examples.

Becker and his collaborators have developed the altruism model in a variety of directions including consideration of conditions under which investments in children will compensate or reenforce ability differences, the role of imperfect capital markets, the determinants of intergenerational income mobility, the effect of altruism on incentives of nonaltruistic family members, and the role of population in economic growth.[4] Others have shown how the empirical implications of the model vary under more general assumptions. For example, Jere Berhman et al. (1982) examine the issue of compensating vs. reenforcing investments in children; B. D. Berheim et al. (1985) show how strategic behavior may influence bequests when children's preferences differ from parents' preferences; and Laurence Kotlikoff and Avia Spivak (1981) demonstrate that intrafamily transfers may substitute for an annuity market in a model without altruism.

The best known implication of the altruism model in macroeconomics is Barro's 1974 "Ricardian Equivalence" result which shows that a public transfer such as pay-as-you-go Social Security will have no effect on real saving, because public intergenerational transfers will be offset dollar for dollar by family intergenerational transfers in the opposite direction. Recently, it has been shown that Ricardian equivalence does not hold when fertility is endogenous because public transfers drive a wedge between the private and social costs of children (D. E. Wildasin, 1985; myself, 1987). For instance, parents have no private incentive to produce children who will pay taxes to support other people in old age; consequently, steady-state fertility will be lower and the capital-labor ratio will be higher in a society with a Social Security transfer program than in one without such a program. My paper (1987) also

[4] See Becker (1981) for a synthesis of and references to his work up to that time and Becker-Tomes (1984) and Becker-Barro (1985) for more recent applications dealing, respectively, with intergenerational mobility and population and economic growth.

points out that this wedge may not occur if Social Security is offset by public intergenerational transfers in the opposite direction such as public schooling; I also show that private fertility choices will be Pareto optimal if there are no market distortions and no net intergenerational transfers through the public sector.

VII. Economic Growth and Demographic Transition

The explanation of the relationship between population and economic growth and development is one of the central challenges to the economics of the family. The theory suggests several possible factors that may be responsible for the demographic transition from high to low fertility which has been associated with the economic development of all the currently advanced societies and in several rapidly growing contemporary developing countries in East Asia and elsewhere.

One explanation of fertility decline, suggested by the "cost of time" hypothesis, is that the increasing value of female time leads to fertility reduction by increasing the relative cost of children. P. Lindert (1980a, b) has shown that the relative cost of children may actually fall during the initial stages of industrialization because the increasing value of child labor offsets the effect of increasing female wages, but documents that an index of child costs does eventually begin to rise as economic development proceeds.

While negative effects of female wages on fertility have been found in many cross-section studies, there are as yet few studies which have attempted to document this effect with historical data. One exception is an interesting and ingenious paper by T. P. Schultz (1985) which uses changes in the relative prices of butter and grain in nineteenth-century Sweden as a natural experiment to test this hypothesis. The basic ideas are 1) that the wage of females relative to males is an increasing function of the relative price of butter to grain because butter was relatively more intensive in female labor than grain; 2) that the relative price of butter to grain is exogenous to labor supply and human capital decisions because these prices were determined in broad international markets; and 3) that there were substantial fluctuations in the relative price. His findings indicate that the rising value of female time accounted for about one-quarter of the decline of Swedish fertility from 1860 to 1910.

A number of factors may influence fertility through quality-quantity interactions in the demand for children which, as explained earlier in the context of the parental altruism model, will tend to be associated with changes in the return to investments in human capital and in the pattern of intergenerational transfers within the family. An influential theory of demographic transition by the demographer, J. C. Caldwell (1976), hypothesizes that the motivation for high fertility in pretransitional societies stems from intergenerational transfers from the younger to the older generation within the family, and that fertility decline is caused by a reversal in the direction of transfers. While Caldwell suggests that the shift in the direction of transfers is the result of increasing individualism caused by modernizing factors such as growth in mass education and the media, I (1982) argued that the underlying cause of both declining fertility and the changing patterns of intergenerational transfers may be improved technology which increases income and raises the return to investment in human capital by shifting the skill and locational distribution of labor demand.

The literature has also considered to varying degrees a wide variety of other factors that may be responsible for fertility decline including decreased mortality, better contraceptive technology, the growth of public transfers and improvements in capital, insurance, and annuity markets.

VIII. Conclusion

What have we learned from the economics of the family? Family economics has just passed its silver anniversary and its father, Gary Becker, has assumed the presidency of the American Economic Association. As a member of this family of economists, I feel pride in its accomplishments, cheered by evidence that it appears to be rising rather than falling, and encouraged by its apparently increasing social acceptability by members of other families within economics and the other social sciences.

But what have we learned? We do not have, as yet, a body of empirically tested, quantitatively stable estimates of the major behavioral relationships suggested by the theory. This state of affairs, unhappily, is not unique to the economics of the family for reasons that are all too familiar and it is not likely to be remedied soon. This, despite the fact that econometric technique in family economics is distinctly state of the art.

We do have a growing capacity to generate hypotheses about both large and small questions concerning family behavior and its consequences within a theoretical framework that is a logically coherent part of the main corpus of neoclassical economic theory. This permits a rich cross fertilization between family economics and other branches of economics that has already borne fruit many times. There is no alternative theory of demographic behavior that comes close in terms of either scope or power. We have a modest body of rejected null hypotheses—yes, economic variables do influence family behavior, often in the direction suggested by the theory—and, as the field has matured and attracted theorists with more varied viewpoints we are beginning to see more sophisticated null hypotheses. Finally, the scope and quality of data on families, both domestically and abroad, has grown enormously during the past twenty-five years. There is even the start of interaction between the theorists and econometricians and the data gatherers. I predict that our family will survive to its golden anniversary.

REFERENCES

Bane, M. J. and Ellwood, D., "Slipping Into and Out of Poverty: The Dynamics of Spells," NBER Working Paper No. 1199, Cambridge, September 1983.

______ **and** ______, "The Impact of AFDC on Family Structure and Living Arrangements," Report to U.S. Department of Health and Human Services, Harvard University, March 1984.

Barro, R. J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–118.

Becker, G. S., "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Universities-National Bureau Conference Series No. 11, Princeton: Princeton University Press, 1960.

______, *Human Capital*, New York: Columbia University Press, 1964.

______, "A Theory of the Allocation of Time," *Economic Journal*, September 1965, *75*, 493–517.

______, *Human Capital and the Personal Distribution of Income*, Ann Arbor: University of Michigan Press, 1967.

______, "A Theory of Social Interaction," *Journal of Political Economy*, November/December 1974, *82*, 1063–93.

______, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

______, "Human Capital, Effort and the Sexual Division of Labor," *Journal of Labor Economics*, January 1985, *3*, S33–S58.

______ **and Barro, R. J.,** "A Reformulation of the Economic Theory of Fertility," Working Paper No. 85-11, Economics Research Center/NORC, October 1985.

______, **Landes, E. M. and Michael, R. T.,** "An Economic Analysis of Marital Instability," *Journal of Political Economy*, December 1977, *85*, 1141–87.

______ **and Lewis, H. G.,** "On the Interaction between the Quantity and Quality of Children," in T. W. Schultz, ed., *Economics of the Family*, 1974, 81–89.

______ **and Tomes, N.,** "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, August 1976, *84*, S142–63.

______ **and** ______, "Human Capital and the Rise and Fall of Families," Working Paper No. 84-10, Economics Research Center/NORC, October 1984.

Ben Porath, Y., "First Generation Effects on Second Generation Fertility," *Demography*, August 1975, *12*, 397–406.

______ **and Welch, F.,** "Do Sex Preferences *Really* Matter?," *Quarterly Journal of Eco-*

nomics, May 1976, *2*, 285–312.

Berger, M. C., "The Effect of Cohort Size on Earnings: A Reexamination of the Evidence," *Journal of Political Economy*, June 1985, *3*, 561–73.

Behrman, J. R., Pollak, R. A. and Taubman, P., "Parental Preferences and Provision for Progeny," *Journal of Political Economy*, February 1982, *90*, 52–73.

Bernheim, B. D., Schleifer, A. and Summers, L., "The Strategic Bequest Motive," *Journal of Political Economy*, December 1985, *93*, 1045–76.

Bernstam, M. S. and Swan, P. L., "The Production of Children as Claims on the State: A Comprehensive Labor Market Approach to Illegitimacy in the United States, 1960–1980," unpublished paper, Hoover Institution, January 1986.

Butz, W. P. and Ward, M. P., "The Emergence of Countercyclical U.S. Fertility," *American Economic Review*, June 1979, *69*, 318–28.

Caldwell, J. C., "Toward a Restatement of Demographic Transition Theory," *Population and Development Review*, September/December 1976, *2*, 321–66.

David, P. A. and Mroz, T. A., "A Sequential Econometric Model of Birth-Spacing Behavior among French Villagers, 1749–1789," Working Paper No. 19, Stanford Project on the History of Fertility Control, Stanford University, January 1986.

De Tray, D. N., "Child Quality and the Demand for Children," in T. W. Schultz, ed., *Economics of the Family*, 1974, 91–116.

Duncan, G. J. and Hoffman, S. D., "A Reconsideration of the Economic Consequences of Marital Dissolution," *Demography*, November 1985, *22*, 485–98.

Easterlin, R. A., "Towards a Socio-Economic Theory of Fertility: A Survey of Recent Research on Economic Factors in American Fertility," in S. J. Behrman et al., eds., *Fertility and Family Planning: A World View*, Ann Arbor: University of Michigan Press, 1969.

______, "Relative Economic Status and the American Fertility Swing," in Eleanor Sheldon, ed., *Family Economic Behavior*, Philadelphia: J. B. Lippincott, 1973.

______, "The Economics and Sociology of Fertility: A Synthesis," in Charles Tilly, ed., *Historical Studies of Changing Fertility*, Princeton: Princeton University Press, 1978.

______, **Pollak, R. A. and Wachter, M. L.**, "Toward a More General Economic Model of Fertility Determination: Endogenous Preferences and Natural Fertility," in R. A. Easterlin ed., *Population and Economic Change in Developing Countries*, Chicago: University of Chicago Press, 1980.

Gronau, R., "The Effect of Children on the Housewife's Value of Time," in T. W. Schultz, ed., *Economics of the Family*, 1974, 457–88.

______, "Leisure, Home Production and Work—The Theory of the Allocation of Time Revisited," *Journal of Political Economy*, December 1977, *85*, 1099–123.

Heckman, J. J., "Effects of Child-Care Programs on Women's Work Effort," in T. W. Schultz, ed., *Economics of the Family*, 1974, 491–518.

______ **and Singer, B.**, "Econometric Duration Analysis," *Journal of Econometrics*, January/February 1984, *24*, 63–132.

______ **and Walker, J. R.**, "The Relationship between Wages and the Timing and Spacing of Births: Evidence from Swedish Longitudinal Data," mimeo., Economics Research Center/NORC, October 1986.

______ **and Willis, R. J.**, "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in N. E. Terleykyj, ed., *Household Production and Consumption*, NBER, New York: Columbia University Press, 1976, 99–138.

Hotz, V. J. and Miller, R., "The Economics of Family Planning," Working Paper No. 85-5, Economics Research Center/NORC, March 1985.

Kotlikoff, L. J. and Spivak, A., "The Family as an Incomplete Annuities Market," *Journal of Political Economy*, April 1981, *89*, 372–91.

Lancaster, K. J., "A New Approach to Consumer Theory," *Journal of Political Economy*, April 1966, *74*, 132–57.

Lee, R. D., "Fluctuations in U.S. Fertility, Age Structure and Income," Final Report to NICHD, Population Studies Center, Ann Arbor, July 1977.

Leibenstein, H., *Economic Backwardness and Economic Growth*, New York: Wiley & Sons, 1957.

Leibowitz, A., "Home Investments in Children," in T. W. Schultz, ed., *Economics of the Family*, 1974, 432–52.

______ **and Eisen, M.,** "An Economic Model of Teenage Pregnancy Decision-Making," *Demography*, February 1986, *23*, 67–77.

Lewis, H. G., "Hours of Work and Hours of Leisure," in *Proceedings of the Industrial Relations Research Association*, Princeton University, 1957, 196–207.

Lindert, P., (1980a) *Fertility and Scarcity in America*, Princeton: Princeton University Press, 1980.

______, (1980b) "Child Costs and Economic Development," in R. A. Easterlin, ed., *Population and Economic Change in Developing Countries*, Chicago: University of Chicago Press, 1980, 5–79.

Michael, R. T., "Education and the Derived Demand for Children," in T. W. Schultz, ed., *Economics of the Family*, 1974, 120–56.

______, "Why Did the U.S. Divorce Rate Double Within a Decade?," in *Research in Population Economics*, Vol. VI, Greenwich: JAI Press, 1986.

______ **and Willis, R. J.,** "Contraception and Fertility: Household Production under Uncertainty," in N. E. Terleykyj, ed., *Household Production and Consumption*, NBER, New York: Columbia University Press, 1975, 27–93.

Mincer, J., "Labor Force Participation of Married Women," in H. Gregg Lewis, ed., *Aspects of Labor Economics*, Universities-National Bureau Conference No. 14, Princeton: Princeton University Press, 1962.

______, "Market Prices, Opportunity Costs, and Income Effects," in C. Christ et al., eds., *Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld*, Stanford: Stanford University Press, 1963.

______ **and Polachek, S.,** "Family Investments in Human Capital: Earnings of Women," in T. W. Schultz, ed., *Economics of the Family*, 1974, 397–429.

Modigliani, F. and Brumberg, R., "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in Kenneth K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick: Rutgers University Press, 1954, 388–436.

Nerlove, M., "Toward a New Theory of Population and Economic Growth," in T. T. W. Schultz, ed., *Economics of the Family*, 1974, 527–45.

Newman, J., "A Stochastic Dynamic Model of Fertility," mimeo., Tulane University, May 1984.

Parsons, Donald O., "The Economics of Intergenerational Control," *Population and Development Review*, March 1984, *10*, 41–54.

Perrin, E. and Sheps, M. C., "Human Reproduction: A Stochastic Process," *Biometrics*, March 1964, *20*, 28–45.

Peters, E., "The Impact of State Divorce Laws on the Marriage Contract: Marriage, Divorce and Marital Property Settlements," *American Economic Review*, June 1986, *76*, 437–54.

Pollak, R. A. and Wachter, M. L., "The Relevance of the Household Production Function and Its Implications for the Allocation of Time," *Journal of Political Economy*, April 1975, *83*, 255–77.

Rosenzweig, M. R. and Schultz, T. P., "The Demand for and Supply of Births: Fertility and its Life Cycle Consequences," *American Economic Review*, December 1985, *75*, 992–1015.

Sanderson, W. C., "On the Two Schools of the Economics of Fertility," *Population and Development Review*, September/December 1976, *2*, 469–78.

______ **and Willis, R. J.,** "Economic Models of Fertility: Some Examples and Implications," in *New Directions in Economic Research*, 51st Annual Report of the NBER, New York, September 1971.

Schultz, T. P., *Economics of Population*, Reading: Addison-Wesley, 1981.

______, "Changing World Prices, Women's Wages, and the Fertility Transition: Sweden, 1860–1910," *Journal of Political Economy*, December 1985, *93*, 1126–49.

Schultz, T. W., *Economics of the Family: Marriage, Children and Human Capital*, NBER, Chicago: University of Chicago Press, 1974.

Stigler, G. J. and Becker, G. S., "De Gustibus

Non Est Disputandum," *American Economic Review*, March 1977, *67*, 76–90.

Weiss, Y. and Willis, R. J., "Children as Collective Goods and Divorce Settlements, *Journal of Labor Economics*, July 1985, *3*, 268–92.

Welch, F., "Effects of Cohort Size on Earnings: The Baby Boom's Babies' Financial Bust," *Journal of Political Economy*, October 1979, *87*, S65–98.

Wildasin, D. E., "Non-Neutrality of Debt with Endogenous Fertility," mimeo., Department of Economics, Indiana University, April 1985.

Willis, R. J., "A New Approach to the Economic Theory of Fertility Behavior," in T. W. Schultz, ed., *Economics of the Family*, 1974, 25–75.

______, "The Direction of Intergenerational Transfers and Demographic Transition: The Caldwell Hypothesis Reexamined," in Y. Ben Porath, ed., *Income Distribution and the Family, A Supplement to Population and Development Review*, 1982, 207–34.

______, "Externalities and Population," in *The Economic Consequences of Population Growth in Economic Development: Contributed Papers*, Washington: National Academy Press, forthcoming, 1987.

Wolpin, K. I., "An Estimable Dynamic Stochastic Model of Fertility and Child Mortality," *Journal of Political Economy*, October 1984, *92*, 852–874.

Multiple Equilibria in Models of Credit

By PETER DIAMOND*

Individuals and firms often want access to considerable amounts of purchasing power on short notice. Production technologies are not generally efficient ways of short-run intertemporal substitution. The sale of nonhomogeneous physical assets is also an inefficient source of flexibility in the access to purchasing power. Rather, the desired purchasing power is generally acquired by the sale of financial assets, use of prior arrangements with financial intermediaries, and borrowing. Use of these institutions frequently involves transaction, information, decision, and monitoring costs. These costs can occur earlier, simultaneously, or later. The presence of these realistic imperfections raises the possibility of multiple equilibria. Technically, lumpiness of transactions can play the same role as transactions costs by ruling out small entries into markets. I review the possibility of multiple equilibria in three models: first, a two-period model of stock market trading; second, a model of bank runs; and third, a model of credit provision in a search setting. There is no claim of underlying unity in these models other than the presence of multiple equilibria. I think of the models as complementary rather than competing.

Central to these models is the assumption of an incomplete set of markets. This incompleteness comes from the same factors that were cited above. However, I do not explicitly model the costs that result in nonexistence of certain trading opportunities. The high costs of certain types of transactions is implicitly given by ruling out that class of transactions. Rather I focus on the costs associated with using the transactions technologies that are assumed to have sufficiently low costs that they appear in the model and are used.

I. Stock Market

Where there is individual uncertainty and an absence of insurance devices, uncertainty arising from planned market transactions depends upon the number of agents entering a market. Thus, thin stock markets, such as the Italian market, are much more volatile than thick ones like the U.S. market. The decision to enter the stock market, both as a shareholder and as a listed firm, depends on the setup costs of going into the market relative to the value of being in. That value depends upon the volatility of the market. Thus, we have the potential for two equilibria: a thin market which has little entry because of volatility and a thick market that has great entry because of lower risks. In *ex ante* terms, the thick equilibrium is superior (and possibly Pareto superior) to the thin equilibrium. Models to capture this idea have been constructed in their Ph.D. dissertations by Satyajit Chatterjee (1986) and Marco Pagano (1986). These models have a first period where participation decisions are made, and a second period where random individual experience generates a distribution of possible market outcomes.

This comparison of thin and thick markets holds when agents behave competitively. The distinction is strengthened when one recognizes the possibility that the potential for market manipulation is probably decreasing with the size of the market over the relevant range of market sizes.

[†] *Discussants*: Ben S. Bernanke, Princeton University; Douglas W. Diamond, University of Chicago; Robert E. Lucas, Jr., University of Chicago.

*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. Research supported by the National Science Foundation under grant no. SES-8509910.

II. Bank Runs

The underlying idea being pursued in the two-period model above is of two different long-run equilibria, either of which could have been reached by a slow-moving historical process. In contrast to this conception, bank runs represent a shift between equilibria which can happen quickly. Bank runs have been modeled by Douglas Diamond and Philip Dybvig (1983) as multiple equilibria in a discrete time model with bank action restricted by limited observation of what is happening within a period. This artificial restriction in a discrete time model is a natural way to capture aspects of a more realistic continuous time model. The critical ingredient here is that the value of an asset portfolio depends on the speed with which it must be sold. This can arise from the cost of evaluation of individual assets (rather than waiting for them to mature), the difficulty of quickly finding eager buyers for idiosyncratic assets, and the market power of hard to find eager buyers. Such problems can arise for individual banks. They can also arise for entire banking systems.

In contrast to the Diamond-Dybvig approach of modeling bank runs as one of (at least) two equilibria, V. Chari and R. Jagannathan (1984) and Andrew Postlewaite and Xavier Vives (1984) have constructed models with a unique equilibrium in which there is a bank run in some states of nature, with accompanying inefficiencies.

The flavor of these results can be brought out in a continuous time Poisson model. Consider an individual with wealth W at time zero. There are two available investments. The illiquid investment opportunity yields a random return r (with mean $\bar{r}$) at time one. If this investment is undertaken, consumption occurs at time one and utility is equal to wealth consumed (i.e., risk neutrality). Thus expected utility from this investment strategy is $W(1+\bar{r})$. Alternatively, the wealth can be held in liquid form giving zero financial return. During time from zero to one, however, there is a flow probability that a "good" consumption opportunity will arrive. If one arrives, it costs precisely W and yields utility $W(1+b)$. The opportunity is only available fleetingly, so that someone without liquid wealth cannot take advantage of it. If no opportunity arrives by time one, the wealth W is consumed. Thus, expected utility for someone holding the liquid investment is $W(F(1+b)+(1-F))=W(1+Fb)$, where F is the probability of the arrival of a consumption opportunity before time one. Assume that the liquid investment strategy is optimal for an isolated individual:

$$\bar{r} < Fb. \tag{1}$$

Now assume two individuals with identical and independent consumption opportunities and perfectly correlated illiquid investment opportunities. Because of the independence of their abilities to profitably use resources early, there is a pooling opportunity. Together, they can hold 0, 1, or 2 times W in liquid form. With full information, total expected utilities for these levels of liquid holdings are then

$$2W(1+\bar{r}),\quad W(2+\bar{r}+(2F-F^2)b),\quad 2W(1+Fb), \tag{2}$$

where $2F-F^2$ is the probability of use of the liquid resources for a good opportunity. In addition to (1), we assume that the aggregate mixed portfolio is optimal:

$$F^2b < \bar{r} < bF. \tag{3}$$

With the mixed portfolio, the two individuals together can do better than they can separately. That is, we assume that pooling of risks raises expected utilities.

We now consider the difficulties of implementing this strategy under various information and observation technologies. We assume that the arrival of a good consumption opportunity is not observable. Remember that such an opportunity is only fleetingly available at some time in the continuous interval between zero and one. The natural symmetric rules are that the first of the pair to try to withdraw the liquid wealth W may do so. The other depositor then receives the illiquid investment at time one. If neither withdraw, they share equally the total wealth available at time one. By our Poisson assumption they don't both attempt to withdraw at the same time. If each one only attempts to withdraw when there is a genuine consumption opportunity, expected utility for each is half the total return from the mixed portfolio with full information. From the assumptions in (3), this is a perfect Nash equilibrium.

There are a variety of ways of creating difficulties for this equilibrium. Following Diamond and Dybvig, we can simply consider the situation if each one begins to think that the other might withdraw funds shortly, whether or not an opportunity arrives. This consideration occurs after funds have been committed to the investments, but before the probability, F, of a good consumption opportunity has decreased. Rushing to withdraw before the other yields $W(1+Fb)$, while waiting yields $W(1+\bar{r})$. Thus belief that the other will definitely withdraw shortly implies that it is optimal to withdraw first. Thus we have two Pareto-comparable Nash equilibria once funds have been committed. The story can be enriched by adding more depositors and allowing assets which can be sold off at a loss. This lowers further the return to late withdrawals in the event of a bank run and reinforces the conditions that give rise to the multiple equilibria. The possibility of a bank run raises the question of whether individuals will commit themselves to this institution at time zero. One can think of the bank run as triggered by a "sunspot." With a sufficiently small probability of this sunspot equilibrium, committing funds to the joint effort is an equilibrium. Diamond and Dybvig argue that deposit insurance has the potential to ease this difficulty.

Following Chari-Jagannathan and Postlewaite-Vives, one can now add signals to the model. These arrive early in time and (for mathematical convenience) not simultaneously to both depositors. Signals could contain information about the random return on the illiquid asset, or about the arrival rates of consumption opportunities. For some signals, immediate attempted withdrawal is the unique Nash equilibrium response. Examples are signals of a sufficiently low illiquid return or sufficiently high probability of either your own "good" consumption opportunity or (particularly perversely) of the other person's. For low probability signals, this can occur as part of a perfect equilibrium with the investment behavior described above.

This contract between these two investors has a clear resemblance to a suspension of convertibility; after one withdrawal, a second withdrawal is refused until the illiquid investment can be realized. Some people have drawn a sharp distinction between suspension and bankruptcy. However, one should recognize that modern bankruptcy law has a similar structure to suspension. A firm or a creditor or regulator of the firm that foresees difficulty in the carrying out of anticipated transactions goes to bankruptcy court for a possible reorganization. This results in a temporary suspension of activities, combined with a revaluation of the claims against the firm. The nature of the conditions thought to be appropriate for triggering suspension and bankruptcy are probably different. Moreover, there are differences in the details of institutional behavior triggered by suspension and bankruptcy. This is particularly seen in the possibility of suspension by many banks which allow interbank transactions but no withdrawals from the banking system.

The bank run model has agents who desire to move rapidly, relative to the speed with which the underlying technology makes liquidity available. Inefficiency in this economy comes from the denial of funds or the costly liquidation which are caused by the bank runs and from the possibility of excessive investment in low yield liquid investments (see, for example, Ben Bernanke and Mark Gertler, 1986). With this interpretation, bank runs seem particularly tied to demand deposits. However, it seems plausible that a model of the cutoff of credit to firms which are regularly rolling over short-term credit would have similar characteristics to the bank run possibility modeled above.

III. Search with Credit

It is natural to ask whether similar multiple equilibria arise when all agents are slow moving; that is, when agents seeking purchasing power move at the same speed as those arranging other transactions in the economy. To answer this question, I have constructed a search model where credit is provided without intermediaries as a possible way of financing pairwise transactions (1986). Moreover, this credit is only provided as part of a transaction in commodities. Thus, action in the credit realm moves at a pace comparable to that in the trade realm. Consider the barter trade model I have analyzed in my 1982 paper. Individuals experience a Poisson arrival of production opportunities. Each opportunity costs c to carry out (a determinate level) and yields a single indivisible unit of good available for trade. Individuals cannot carry more than one unit of the tradable good in inventory. There is a second Poisson process which brings together individuals randomly selected pairwise in the economy. A fraction of these meetings results in trade when both individuals have goods available for trade, or when one of them does and is willing to provide credit to the other one. To make the most sense of the underlying model, one should think of the problem as being that of finding goods to one's taste rather than that of finding goods at all. Secondly, assume that the desire for variety implies that one never wants to have the good produced by a given individual more than once in one's lifetime. Thus there is no potential for long-term arrangements between two individuals.

When individuals engage in a barter trade, the symmetry of their positions implies that they exchange the goods one for one. When they engage in a trade with credit, a unit of good in inventory is traded for the next unit to be produced by an agent without inventory. This trade would have an interest rate of zero. One can have a smoothly varying implicit interest rate by assuming that the debtor provides his next produced unit of tradable good to the lender with probability 1, but the lender only delivers the good he has on hand with probability p. The probability of delivery p is chosen to satisfy the Nash bargaining solution for the trade opportunity between this pair of individuals. To look for multiple equilibria, the question is asked whether there are parameters for the underlying economy such that a credit limit of zero (no trade on credit) and a credit limit of one unit of goods are both equilibria.

In the Arrow-Debreu model, the credit limit for an individual is determined by the ability to pay back the debt over the rest of life. In models with monitoring costs, this credit limit is adjusted for the stochastic nature of the ability to repay debt and the costs of monitoring and collecting arising under uncertainty. Once one recognizes that individuals can choose not to pay back, one must identify the penalties for such a choice and have a credit limit reflecting incentives to pay back as well as an inability to pay back. In order to model the credit limit in this way, I have assumed that the alternative to delivering the next produced good to the creditor is to drop out of the trading network. The expected utility of dropping out is taken to be exogenous and used as the origin. Thus the individual compares the shadow value of being able to continue trading in the economy with the cost of carrying out a project and delivering it to the creditor. Contrasting two possible equilibria in the economy with credit limits of zero and one, there are three differences. One is that a greater credit limit makes it more valuable to be in the trading economy at any level of debt. Second, the probability of receiving goods in a credit transaction, that is, the implicit interest rate in the transaction above, depends upon the credit limit. In the examples I have calculated, there is a lower implicit interest rate in the economy with a larger credit limit. Third, the availability of credit feeds back on the stock of inventories in the economy since a greater level of trade implies a lower level of the stock of inventory. (It also implies a higher level of production since more individuals are available to carry out production opportunities.) The model described above has multiple equilibria for a large fraction of the parameters for which there exists an equilibrium with a credit limit of zero. Thus it appears that the positive feedback loops which generate multiple equilibria in credit models do not require fast moving agents to have multiple equilibria.

Since the pairwise trade transactions are lumpy, credit provision in this model is also lumpy. This lumpiness plays a central role in the modeling of multiple equilibria. While the arrangement of credit for lumpy projects is indeed lumpy, the possibility of smooth accumulation of resources before organizing credit smooths one aspect of credit arrangements. This smoothing is missing in the model.

There are two reasons for my being interested in the possibility of multiple equilibria as a phenomenon associated with credit transactions. One is that the presence of multiple equilibria in a fast-moving setting may be the natural way to model the role of a lender of last resort as well as being an important element to be considered in the design of institutions in the credit market. In this way one can recognize that central banks can move very quickly. Second, the presence of a dynamic model with two different steady-state equilibria introduces the possibility that the dynamic economy may at some times be near the borderline between the capture areas of the two different equilibria under the appropriate adjustment dynamics. If this happens, relatively small actions by a government can induce substantial differences in the long-run equilibrium position of the economy. One could make sense of the idea of pump priming in this way.

Since it is precisely the frictions that make credit an interesting question (and the Modigliani-Miller Theorem inapplicable) it seems necessary to construct equilibrium models with frictions to make sense of both short-run and (institutional) long-run policy questions for central banks and bank regulators. In a richer model one would like to analyze the structure of financial intermediaries and the properties associated with different equilibrium structures. It seems likely that there will be multiple equilibria in structures (even in the absence of government regulation) with little claim to optimality for the equilibrium structure given the nature of externalities arising with credit provision. This rich research agenda is in its infancy. It looks to be an extremely interesting area in which to work and one with sizable potential payoffs for policy.

REFERENCES

Bernanke, Ben and Gertler, Mark, "Banking and Macroeconomic Equilibrium," unpublished, February 1986.

Chari, V. and Jagannathan, R., "Banking Panics, Information and Rational Expectations Equilibrium," BRC Working Paper No. 112, Northwestern University, J. L. Kellogg School of Management, 1984.

Chatterjee, Satyajit, "Market Participation and Macroeconomic Equilibrium," working paper, University of Iowa, June 1986.

Diamond, Douglas W. and Dybvig, Philip H., "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, June 1983, *91*, 401–19.

Diamond, Peter, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–894.

______, "Stochastic Credit in Search Equilibrium," Working Paper No. 421, MIT, June 1986.

Pagano, Marco, "Endogenous Stock Market Size and Asset Price Volatility," working paper, MIT, March 1986.

Postlewaite, Andrew and Vives, Xavier, "Bank Runs as an Equilibrium Phenomenon," unpublished, University of Pennsylvania, June 1984.

Ultimate Sources of Aggregate Variability

By ROBERT J. SHILLER*

Any empirical model of the macroeconomy tells a story about the exogenous shocks that are ultimately responsible for changes from year to year in macroeconomic variables. The characterization of these sources of aggregate variability is of fundamental importance. Economic theory cannot be applied to data unless we know which economic relations are *not* themselves shocked, or at least unless we know something about the shocks.[1]

Recent models have differed widely in their characterizations of these ultimate sources of aggregate variability. Finn Kydland and Edward Prescott (1982) and others proposed models of the business cycle in which the *only* shocks to the macroeconomy are certain kinds of shocks to technology. Robert Barro (1977) proposed a model in which 78 percent of the variance of a (transformed) U.S. unemployment rate 1946–73 is due to unexpected changes in the money stock, and military employment and minimum wage variables. David Lilien (1982) argued that most of the unemployment fluctuations in the United States in the 1970's were due to "unusual structural shifts," such as changes in the demand for produced goods relative to services. James Hamilton (1983) argued that dramatic oil price shocks preceded all but one of the recessions in the United States since World War II, and that these oil price shocks were in turn caused by events such as the nationalization of Iranian assets, the Suez crisis, and strikes by oil and coal workers.

Other have offered analyses that, while not necessarily claiming to isolate the major source of aggregate fluctuations, do suggest that qualitatively very different exogenous shocks may be quite important: changes in desired consumption (Robert Hall, 1986), breakdowns in the process of borrowing and lending (Ben Bernanke, 1981), breakdowns or establishments of cartels (Julio Rotemberg and Garth Saloner, 1986), or variations in attitudes toward union membership (Olivier Blanchard and Lawrence Summers, 1986). Moreover, any of these shocks might occur in a foreign country, and be transmitted by trade relations to the domestic economy. Some analyses have even emphasized that something that ought to be, by any fundamental logic, truly irrelevant to the macroeconomy, may well importantly influence it if people think it does (for example, Michael Woodford, 1987). The potential importance of such variables may be even more important than suggested in some papers in the theory literature if we allow for "near-rational expectations" as well as the strictly rational expectations.

Recent evidence (John Campbell and N. Gregory Mankiw, 1987) suggests that innovations in real gross national product (GNP) show little tendency to be reversed subsequently, and that the apparent tendency of GNP to be trend-reverting may be due to spurious trend estimation. To the extent that this is right, then variations in the same sources that explain long-run growth (and explain why the United States is wealthier than India or China) would also play a role in explaining short-run movements. If cultural or institutional factors influencing the dissemination and application of learning are the reason, then changes in these factors may play a role. If economies of scale are a factor determining intercountry differences, then the discovery of new industries or regions of production functions where such economies obtain might also play a role. Other possible factors are changes in government expenditure on "infrastructure"

*Yale University, Cowles Foundation for Research, New Haven, CT 06520. The research was supported by the National Science Foundation. Ray Fair was very helpful in framing this paper.

[1] Peter Garber and Robert King (1983) pointed out that contemporary Euler equation estimation methods always assume that the shocks come in somewhere else in the model, but that this assumption will not do for every equation in the model.

formation, government policies and other factors encouraging or discouraging initiative, or just population growth and natural resource discoveries and depletion.

There seem to be a bewildering array of possibilities for ultimate sources. It is important to understand that it is in principle possible that they might all contribute substantially. It is technically possible that 10 different independent shocks each might make a contribution whose standard deviation is 32 percent ($=\sqrt{0.1}$) of the standard deviation of the aggregate. Thus, for each of the 10 factors there may be evidence that it is often very important and occasionally dominates aggregate fluctuations. Or, it is possible that 100 independent factors each may make a contribution whose standard deviation is 10 percent of the aggregate standard deviation.

One is naturally led to wonder if there isn't any systematic way to determine what is the relative importance of different sources of macroeconomic variability.

I. Pigou's Analysis of Ultimate Sources

A. C. Pigou's remarkable book *Industrial Fluctuations* (1929) appears to be the most recent effort until now to provide such a systematic breakdown of sources. He grouped these sources into three broad categories. "Real Causes" are "changes that have occurred, or are about to occur, in actual industrial conditions and expectations based on these are true or valid expectations." The principal real causes he cites are: 1) harvest variations, 2) inventions, 3) industrial disputes, 4) changes in fashion, 5) wars, and 6) foreign demand and foreign openings for investment. "Psychological causes" are "changes that occur in men's attitude of mind, so that, on a constant basis of fact, they do not form a constant judgment." "Autonomous monetary causes" are events affecting money, such as gold discoveries, or changes in monetary or banking policies (p. 35). He thought that removal of either the autonomous monetary or the psychological causes might reduce the amplitude of industrial fluctuations by about a half. Removal of harvest variations might reduce amplitude by about a quarter. He thought that other real causes, such as inventions or work stoppages, had much less effect.[2]

Of the above sources, only one, the psychological, appears largely absent from contemporary macroeconomics, though it might be interpreted as present in some macroeconometric models in the form of error terms. Pigou here describes swings in optimism or pessimism affecting investment that arise "spontaneously," though perhaps ultimately as a "psychological reflex" from some of the same factors that he calls "real" causes (p. 73). He emphasized that the swings occur simultaneously over a large number of people, because of "psychological interdependence," "sympathetic or epidemic excitement," or "mutual suggestion" (p. 86).[3] He denied what we now call "rational expectations" because there is "instability in the facts being assumed," though he admitted that "if everything were absolutely stable, recurring every year with exact similarity or in a perfectly regular progression, people could not fail to be aware of the relevant facts and to form correct judgments" (p. 74).

What sort of evidence might Pigou have that his list comprises the important sources? Although he made some use of statistics, his method involves judgment that appears to be based on anecdotal and narrative historical evidence. Such a method may in fact be of some use for this purpose. A rough sense of proportion about some economic mechanisms may suggest that if certain factors would change exogenously, there would be important macroeconomic consequences. If these factors did indeed change historically, then the only question that remains is whether such a change could be purely endogenous, that is, caused ultimately by other economic variables. We may then have some idea whether these factors are likely to be determined reliably by such other economic variables.

[2] See pp. 219–25. Note that Pigou's breakdown denies independence of factors: he thought eliminating one may reduce the impact of another.

[3] On this point, compare Woodford.

Consider, for example, the case of the autonomous monetary causes. In the contemporary context, we know that the Fed can and does move interest rates, which have a major impact on the economy. Of course, the announced goal of the Fed currently is to stabilize the economy, and their efforts may indeed attenuate the effects of other shocks. But since their methods are judgmental and imprecise, it is to be expected that they must also serve to *add* shocks themselves. By analogy, if we analyzed the movements of an airplane in rough weather, we would expect to find that a component of the airplane's movements is ultimately due to the pilot.

Methods like Pigou's are suggestive, but one might hope for something more objective and quantitative.

II. Evidence from Large-Scale Macroeconometric Models

The large-scale macroeconometric models in the Keynesian tradition appear to be the only models detailed enough to allow a decomposition of output variability into a variety of constituent shocks as broad as that proposed by Pigou. In these models, all macroeconomic fluctuations can be traced ultimately to equation residuals or exogenous variables.

Ray Fair (1986) has undertaken stochastic simulations of Fair model of the U.S. economy to show what are the important shocks to the model. The Fair model is similar to most large-scale macroeconometric models in that it includes consumption and investment functions, and a national income identity to yield an IS curve, and demand for money equations that gives rise to an LM curve. Monetary policy is modeled by a Fed reaction function, but fiscal policy is taken to be exogenous.

To take account of shocks to exogenous variables, he added simple autoregressive forecasting equations for 23 exogenous variables to the 30 structural equations in his model, producing a 53-equation model with basically no exogenous variables. Taking as given data through 1981:II, a stochastic simulation, the base simulation, was run using a 53×53 (block diagonal, with a 30×30 and 23×23 block) variance covariance matrix residuals, and the variance of actual real GNP for one to eight quarters ahead (i.e., 1981:III through 1983:II). He then set residuals for the eight quarters to zero in each of the 53 equations, one at a time and then in groups, and ran new stochastic simulations. The variance in real GNP in any one of these simulations as a percent of the variance in the base simulation is a measure of the importance of the residual that is analogous to the square of the relative amplitudes described by Pigou.

What is striking about the results is that the conclusions differ substantially between one-quarter-ahead simulations and eight-quarter-ahead simulations. For example, if we drop the error term to the inventory investment equation, real GNP variance falls by 29 percent relative to the base simulation in the one-quarter-ahead simulations, but by only 4 percent in the eight-quarter-ahead simulations. If we drop the error term in all investment equations (consumer durables, housing, inventories, and business fixed investment) the corresponding figures are 50.6 and 13.4 percent. Thus, failure to predict investment accounts for most of the model's difficulty in forecasting one quarter ahead, but relatively little of the difficulty in making longer-run forecasts. Other sources of variability grow faster with time horizon, so that uncertainty about investment is swamped out.

The story told by the Fair model is a complicated one, with no single source of variability dominating. Consider the percentage variance declines in the eight-quarter-ahead simulations for real GNP. Dropping all exogenous variables' shocks reduced variance by 44 percent, the remainder being accounted for by equation residuals. The principal grouping of exogenous variables was government expenditure and transfers (federal, state, and local), for 21 percent and after that, exports, for 19 percent. Among endogenous variables, dropping residuals on consumption on services and nondurables reduced variance by 10 percent, on the wage and price sector by 11 percent, and on import demand by 7 percent. Dropping Federal Reserve policy shocks reduced variance by only 3 percent.

What sort of evidence is behind the Fair Model that gave rise to these variance decompositions? The modeling effort relied on the assumption that a large list of variables is exogenous. Many of these are not plainly exogenous to the model, though one might suppose that their relation to economic activity is in some cases tenuous, complicated, and involving long lags. The modeling effort also relied on a set of restrictions on coefficients that vastly overidentified the model. Because of these overidentifying restrictions, the estimate of the reduced form was not at all the same as if it had been estimated by merely regressing endogenous variables on exogenous and predetermined variables. The restrictions sometimes have the effect, for example, of inferring an effect of an exogenous variable on GNP from an observed effect on a component of GNP.

These overidentifying restrictions were usually not explicitly discussed in the description of the model. Their specification appears to have largely intuitive origins, just as was the case with the theory of Pigou. One would wish that there were a method that was more capable of producing a consensus in the profession as to ultimate sources.

III. Partial Specifications of Exogenous Sources

Many doubt the assumptions of the large-scale macroeconometric models. But, certainly *some* of their assumptions must be uncontroversial. Certainly *some* variables (for example, the weather) would be judged genuinely exogenous by just about everyone. It would be progress if we could all agree that such a variable explains x percent of macroeconomic variability, even if x is very small. Might not a Granger or Sims causality test for causality from such a variable to real GNP produce such an agreement?

But it's hard to think of any single measurable clearly exogenous variable that seems likely to have much impact on the aggregate economy. We have a wealth of data at a finely disaggregated level, for example, information on individual patents each of which represents a component of technological progress. But how to aggregate this information into a data series that might be found to cause GNP? We cannot regress GNP on hundreds of exogenous variables each of which explains a component of it, since we would have more independent variables than available observations.

Weather variables are probably the most obvious candidates for a truly exogenous variable that might really cause macroeconomic aggregates. Regression models explaining individual crop yields (for example, Wolfgang Baier, 1977) show that weather variables explain a substantial portion of year-to-year crop variability. Often the R^2 is over 0.5. But to achieve such R^2 for individual crops the researcher uses finely focused weather variables that differ across crops, such variables as "estimated June potential evapotranspiration," or "mean soil moisture reserves (mm) at heading stage in 0–100 cm. depth of soil." To explain aggregates well, these weather variables should be measured at all the appropriate times and sites for the specific crops to which they pertain. To explain weather effects on nonagricultural productive activities would require yet very different weather variables. To explain housing starts, we may use number of days in the year where temperature is below freezing, to explain restaurant meals the number of evenings of inclement weather and highway conditions in urban areas, to explain electricity demand an average of a nonlinear function of summer temperatures above 75 °, to explain heating fuel demand an average of a nonlinear function of winter temperatures below 60 °.[4] It is not easy to find a good aggregator of these shocks other than GNP (or its analogues) itself.

[4] Donald Deere and Jeffrey Miron (1986) regressed U.S. layoff rates by state and industry on state-specific (but not finely focused) weather variables and other variables. The weather variables were significant at the 90 percent level in about 25 percent of the regressions, and were very significant overall.

IV. Models as Aggregators

Finding an aggregator of very many exogenous shocks means building a highly disaggregated model that explains many components of GNP and shows how they interact to produce the total. If models are to be judged as aggregators, then models may be deemed successful even if they have known structural defects that would cause them to be rejected by conventional criteria. An aggregator model might be only a naive or crude model. For example, we might build a large Leontief input-output model. Data on the implementation of technological innovations, on weather, or on other known exogenous shocks could be used to adjust the elements of the input-output matrix and the matrix of factor-input requirements. We might find that an index of structural change in the model aggregates successfully (for example, Granger-causes GNP) even though we know that the assumption of fixed proportions is highly restrictive.

Large macroeconomictric model projects that deal laboriously with details may thus yield insight into sources of variability. Existing large-scale macroeconometric models may be viewed, even by those who accept some of the well-known criticism of their theory, as having shown some such success already (see Fair and myself, 1987). It is natural to expect that further progress can be made along these lines, taking account of developments in economic theory and data, if people are willing to do more work at a detailed level, and for many different countries.

V. Interpretation

There is as yet no consensus in the profession as to the quantitative importance of *any* of the various ultimate sources. In my judgment, however, the existing literature does suggest that a great multiplicity of sources is at work: shocks to tastes as well as technology, shocks in government policy, demographic shocks, shocks to organizations in labor or industry, and "psychological" shocks of the kind described by Pigou and others.

Currently popular methodology results in models that attempt to make do with very few shocks. These models are valuable as special cases but should be interpreted as exploratory exercises. We should not consider it an objective of research to simplify or reduce the array of exogenous shocks. Simplicity is of course a virtue, but simple models cannot be construed as an objective if the world is not simple.

REFERENCES

Baier, Wolfgang, "Crop-Weather Models and their Use in Yield Assessments," Technical Note No. 151, World Meteorological Organization, 1977.

Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

Bernanke, Ben S., "Bankruptcy, Liquidity, and Recession," *American Economic Review Proceedings*, May 1981, *71*, 155–59.

Blanchard, Olivier and Summers, Lawrence, "Hysteresis and the European Unemployment Problem," reproduced, MIT, February 1986.

Campbell, John Y. and Mankiw, N. Gregory, "Permanent and Transitory Components in Macroeconomic Fluctuations," *American Economic Review Proceedings*, May 1987, *77*, 111–17.

Deere, Donald R. and Miron, Jeffrey A., "The Cross Sectional Impact of Unemployment Insurance on Layoff, Employment and Wages," reproduced, Texas A&M University, 1986.

Fair, Ray C., "Sources of Output and Price Variability in a Macroeconometric Model," reproduced, Yale University, 1986.

______ **and Shiller, Robert J.,** "Econometric Modelling as Information Aggregation," reproduced, Yale University, 1987.

Garber, Peter M. and King, Robert G., "Deep Structural Excavation? A Critique of Euler Equation Methods," NBER Technical Working Paper No. 31, November 1983.

Hall, Robert E., The Role of Consumption in Economic Fluctuations," in Robert J. Gordon, ed., *The American Business Cycle:*

Continuity and Change, NBER, Chicago: University of Chicago Press, 1986.

Hamilton, James D., "Oil and the Macroeconomy since World War II," *Journal of Political Economy*, April 1983, *91*, 228–48.

Kydland, Finn E. and Prescott, Edward C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Lilien, David M., "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *90*, 777–93.

Pigou, Arthur C., *Industrial Fluctuations*, 2nd. ed., London: Macmillan, 1929.

Rotemberg, Julio J. and Saloner, Garth, "A Supergame-Theoretic Model of Price Wars during Booms," *American Economic Review*, June 1986, *76*, 390–407.

Woodford, Michael, "Self-Fulfilling Expectations and Business Cycles," *American Economic Review Proceedings*, May 1987, *77*, 93–98.

Three Questions about Sunspot Equilibria as an Explanation of Economic Fluctuations

By MICHAEL WOODFORD*

It is by now well known that the sort of difference equations that characterize the equilibrium conditions of an infinite horizon competitive economy may have solutions in which the endogenous variables fluctuate in response to "sunspot" variables, that is, to random events that in fact have nothing to do with economic "fundamentals," and so do not directly affect the equilibrium conditions. It is possible to view such "sunspot equilibria" as a representation of an actual phenomenon—economic fluctuations not caused by exogenous shocks to fundamentals, but rather by revisions of agents' expectations in response to some event, which revised expectations become self-fulfilling.

Early discussions of such solutions sometimes suggested that a more rigorous derivation of the requirements for equilibrium might yield additional restrictions that would eliminate the sunspot solutions from the set of true equilibria. The demonstration by Karl Shell (1977), David Cass (1981), and Costas Azariadis (1981) that sunspot equilibria can exist in a rigorously formulated intertemporal equilibrium model, namely the overlapping generations model of Samuelson, has shown that this is not always the case. Nevertheless, many economists remain skeptical about the reasonableness of the sunspot hypothesis as a possible explanation of actual economic fluctuations, and for quite general reasons, independent of judgments about the empirical plausibility of any particular models. I discuss here three such general reasons for skepticism.

I. Are there Unexploited Profit Opportunities for Market Makers?

One argument against the practical significance of sunspot equilibria is that their existence is often linked to market incompleteness. It may be argued that, while the absence of markets for claims contingent upon sunspot realizations is not surprising in a world in which such events have no economic significance, if fluctuations of any size in response to sunspot realizations were to occur, this would provide an incentive for someone to organize such markets, and so suppress the fluctuating equilibria.

A complete set of Arrow-Debreu markets would indeed rule out sunspot equilibria, under quite general assumptions (Yves Balasko, 1983). For example, in an exchange economy of the kind considered by Shell, Cass, and Azariadis, if a sunspot equilibrium exists, an allocation that would give each agent in some period t the mean of the vector of goods received in the sunspot allocation, and that coincides with the sunspot allocation in all other periods, must also be feasible. If agents are risk averse, the alternative allocation increases the expected utility of all agents who consume in period t, while not affecting that of any other agents. But then at any set of Arrow-Debreu prices, the mean allocation for at least one agent must cost no more than his sunspot allocation, and so the sunspot allocation cannot be an Arrow-Debreu equilibrium. (The example of Cass, showing that sunspot equilibria are possible in an overlapping generations exchange economy with complete contingent claims markets, depends upon all agents having linear utility functions.)

One answer to the objection that markets should open in the event that sunspot fluctuations occur is given by Azariadis: there are very many possible random variables that could play the role of sunspots, and if markets are opened to neutralize certain of these variables, equilibrium fluctuations may be coordinated instead by another variable for which there are no contingent claims markets. But this is not particularly satisfactory, since (as discussed below) there must

*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637.

be some reason for all agents to simultaneously come to expect a particular sunspot variable to matter, and one may well suppose that anything that serves to render a particular variable sufficiently salient for this purpose can equally well predispose financial innovators to establish a market for exactly that sort of contingent claim.

A more satisfactory reply would observe that the kind of trading in contingent claims that rules out the sunspot equilibria in a model like that of Azariadis involves agents insuring against the risk of being born in an undesirable sunspot state. Suppose that one opens markets for all contingent claims, but only allows trading by agents *after* they have been born (and after revelation of the sunspot state in their first period of life). In this case sunspot equilibria are still possible in the Azariadis model; indeed, all of the equilibria considered by Azariadis still exist. One needs only to work out a set of contingent claims prices at which no agents who are allowed to wish to trade these claims (starting from one of Azariadis' sunspot allocation). This is easily done, since in the case of any two contingent claims, there is at most one type of agent who consumes in both states. Thus sunspot equilibria are consistent with trading in contingent claims *after birth*, and this kind of trading is the only kind whose absence could represent an unexploited profit opportunity in a model of this type.

Another kind of economy in which sunspot equilibria may exist even in the case of trading in claims contingent upon sunspot histories is a cash-in-advance monetary economy of the sort studied by Robert Lucas and Nancy Stokey (1987). Suppose that an infinite lived representative agent seeks to maximize the expected value of

$$\sum_{t=1}^{\infty} \beta^{t-1}[u(c_t)-v(n_t)],$$

where n_t is output supplied for sale in period t, and c_t is goods purchased in period t. Goods purchases are subject to a cash-in-advance constraint, so that income from supply of output can only be spent in the following period. Let us suppose that the money supply is increased at a constant rate g (the per capita money supply in period $t+1$ is g times that in period t), and that additions to the money supply each period are through lump sum transfers (or taxes if $g<1$) that add to (or subtract from) the money balances from which agents can finance goods purchases in that period. Now let $\hat{z}$ denote the solution to $u'(\hat{z})=v'(\hat{z})$, that is, the level of per capita real balances at which the cash-in-advance constraint ceases to bind. Then a rational expectations equilibrium is a stochastic process for per capita real balances z_t that satisfies

$$(1)\qquad F(z_t)=(\beta/g)E_t[G(z_{t+1})]$$

and a transversality constraint, where

$$F(z)\equiv zv'(\min(\hat{z},z)),$$
$$G(z)\equiv zu'(\min(\hat{z},z)).$$

Given any such process for z_t, the equilibrium allocation of resources is given by

$$c_t=n_t=\min(\hat{z},z_t).$$

Difference equation (1) is of the same form as the equilibrium condition considered by Azariadis, and the methods of that paper can be used to show that stationary sunspot equilibria are possible in a Lucas-Stokey economy. (Any finite-state Markov process solution to (1) necessarily satisfies the transversality condition as well; see my paper 1986c.) In fact, in the case of equilibria in which the cash-in-advance constraint always binds ($z_t\leq\hat{z}$ always), the equilibria of the Lucas-Stokey model just described are identical to those of an overlapping generations model of the Azariadis type in which agents born in period t seek to maximize $E_t[\beta u(c_{t+1})-v(n_t)]$.

The existence of sunspot equilibria in this model depends upon market incompleteness, because of the argument given above; but it does *not* depend upon the absence of trading in claims contingent upon sunspot histories. In fact, one can introduce markets for such claims without changing the set of equilibria at all. Following Lucas and Stokey, let each period be divided into two subperiods. In the first subperiod, financial markets are open, in which spot money is traded against obligations to pay money in the future, which may be contingent upon future sunspot realizations. In the second subperiod, goods markets are open, in which goods may be purchased using money held at the end of the first subperiod. We suppose that the sunspot state is revealed at the beginning of the first subperiod. Lucas and Stokey show that the equilibria of this model with a complete set of contingent claims markets of this sort are identical to those of the model with no financial markets described above. (They do not explicitly consider sunspot states, but their analysis applies equally to this case.)

Thus it is *the cash-in-advançe constraint* that plays the crucial role in allowing sunspot equilibria in this model, not any absence of trading in claims contingent upon sunspot realizations. Of course, it is appropriate to scrutinize the microeconomic foundations of the cash-in-advance constraint, and ask why it is that no agent has an incentive to engage in financial innovation that would relax that constraint. But there is no reason to expect that the existence of sunspot fluctuations of a substantial magnitude should greatly *change* such incentives from what they are in the case of, say, the quantity-theoretic equilibrium in which prices increase at the constant rate g. Accordingly, it does not seem that the existence of sunspot equilibria must necessarily involve any opportunity to profit through opening new markets, that should be expected to suppress such equilibria.

II. Can Sunspot Theories Yield Useful Predictions?

Another general objection to sunspot equilibria as positive models of economic fluctuations would argue that models in which equilibrium is indeterminate (as is true of all sunspot models) are by that virtue models that yield few definite predictions. It might also be argued that insofar as it is impossible to perform comparative statics analysis of the effects of policy interventions with a model in which equilibrium is indeterminate, acceptance of such models means abandonment of any hope of using economic theory as a guide for policymaking. Accordingly, some would reject sunspot equilibria as positive models in all cases. Either, they would argue, one should only propose as candidate models of the economy models in which equilibrium is determinate; or one should try to reduce the set of equilibria in order to sharpen the prediction of one's model, by, for example, ruling out all equilibria in which sunspot variables matter.

There are a number of possible responses to this objection. First, the mere fact that one model has a unique equilibrium does not mean that it yields sharper predictions than another model with multiple equilibria; it depends upon how many "unobserved shock" terms are postulated by the first model. (See my papers, 1986a; 1987, sec. 4, for demonstrations that sunspot models can yield many quite specific predictions regarding the character of equilibrium fluctuations.) If fluctuations were to be explained as entirely due to shifts in observed exogenous variables in a model with a determinate equilibrium, exact comovements of the various variables of the model would be implied, of a kind that are in fact not observed. Hence all econometric models allow for a large number of unobserved shocks to various equations of the model; and in traditional macroeconometric models, such equation residuals account for more of the variability of the endogenous variables than do shifts in observed exogenous variables. Replacing simple equation residuals by explicit shocks to tastes and the like does not greatly increase the extent to which such models can be said to truly explain economic fluctuations, since these shocks only fulfill their function of rendering nonsingular the predicted covariance matrix of the data insofar as the existence of the shocks cannot be independently observed. A sunspot theory that predicts a nonsingular covariance matrix (under certain conditions, whose validity may itself be tested) without the need to postulate such shocks might well be considered a less *ad hoc* "theory of the error term."

Moreover, it may be argued that the very fact that equilibrium is indeterminate in certain economic models, so that revisions of expectations may be self-fulfilling, is itself a prediction of great importance; the fact that *sunspot equilibria exist in certain models and not in others* allows one to make predictions about the degree of instability that should be associated with different economic structures, assuming similar levels of variability in "fundamentals." Predictions of this kind could be tested by comparing the experience of different countries and different historical periods with differing regulatory and policy regimes. For example, many recent applications of the concept of sunspot equilibrium draw attention to the role of certain types of financial institutions in allowing equilibrium fluctuations of this type. Douglas Diamond and Philip Dybvig (1983) link the possibility of bank runs due to purely extrinsic uncertainty to particular features of demand deposit contracts, and show how particular institutional arrangements, such as deposit insurance or suspension of convertibility in certain circumstances, can suppress such equilibria. Bruce Smith (1986) exhibits an economy in which sunspot fluctuations are possible in the case of free banking, but can be suppressed by regulations that restrict inside money creation, and compares his results to nineteenth-century British debates about banking instability. My papers (1986a, 1987) show that cyclic fluctuations in investment may occur as a sunspot phenomenon in an economy in which restrictions upon the ability of wage earners to borrow against future wage income coexist with an elastic supply of bank loans to entrepreneurs; either an improvement in the availability of consumer credit or control of the nominal volume of bank loans to entrepreneurs can suppress such fluctuations. Results of this kind provide a potential basis for interesting historical and international comparisons.

Furthermore, predictions of this kind may be useful for policy analysis insofar as rendering equilibrium determinate and suppressing sunspot fluctuations may itself be an object of policy. Determinacy of equilibrium is then not necessary in order for economic theory to be useful for the evaluation and design of public policy; and *assuming away* multiple equilibria, arguing that the sunspot equilibria are not "economically relevant," may mean throwing away the prediction of one's model that is of greatest importance.

As an example of what may be missed in policy analysis that ignores the existence of sunspot equilibria, consider again the Lucas-Stokey model, and suppose that the monetary authority contemplates a change in the rate of growth of the money supply. To each value of g between β and some (possibly infinite) upper bound, there corresponds a unique quantity-theoretic equilibrium in which prices grow at the constant rate g and output never varies. If one simply compares the welfare of the representative agent across these deterministic steady states, one finds that welfare is monotonically decreasing in g, so that one should seek to reduce g to be as close to β as possible (Friedman's "optimum quantity of money").

On the other hand, as noted above, there may exist sunspot equilibria in this model, and changing g can affect whether the economy is unstable in this sense. (Let us assume that the government has adopted policies of the sort discussed by W. A. Brock and J. A. Scheinkman, 1980, in order to rule out equilibria in which per capita real balances ever become extremely large or extremely small; in the absence of such policies, for many kinds of preferences it turns out that hyperinflationary sunspot equilibria exist for all rates of money growth consistent with the existence of any monetary equilibrium, for the reason discussed by those authors.) One often finds (see my 1986c paper) that sunspot equilibria exist for low, rather than high, values of g—exactly the values of g that would be considered most desirable if only the welfare properties of the steady state are considered. The intuition is simple. A condition like (1) has sunspot solutions in which z_t remains forever bounded above and away from zero only if a change in the expected value of z_{t+1} has sufficiently strong effect upon the equilibrium value of z_t; and for any given slopes for the F and G functions, a *lower* value of g means a *larger* effect of any given change in expectations regarding z_{t+1} upon the right-hand side of (1), and hence a larger change in z_t is required to reestablish equilibrium. Thus a lower value of g strengthens the destabilizing positive feedback loop from expectations of price level volatility to actual price level volatility.

My 1986c paper also shows that the expected utility of the representative agent in a stationary sunspot equilibrium associated with a low value of g can be *lower* than the utility obtained in the unique equilibrium (the quantity-theoretic steady state) associated with a higher value of g, even though the level of utility in the steady state associated with the low value of g is *higher*. Hence if one were to assume that when g is low enough for nonexplosive sunspot equilibria to exist, the economy will in fact move to an equilibrium of this sort, then decreasing g to this point would decrease the welfare of the representative agent. If one were to suppose that when rational expectations equilibrium ceases to be unique, agents are unlikely to be able to coordinate their expectations upon any equilibrium at all, the consequences of too low a rate of money growth could be even worse.

III. Could Agents Ever Come to Have such Expectations?

If, however, the existence of sunspot equilibria as a theoretical possibility gave one no reason to believe that agents would be any less likely to coordinate their expectations upon one of the nonsunspot equilibria, there would be no reason to design institutional structures or stabilization policies so as to suppress the sunspot equilibria. A final general objection to sunspot equilibria as a positive model of economic fluctuations argues that it is implausible that agents should ever come to have the expectations associated with a sunspot equilibrium. For it becomes rational for agents to respond at all to the realizations of a sunspot variable only after a large part of the population already responds to that variable, and all in the same way. One may wonder how such a coherent pattern of response could ever get started.

My 1986b paper explicitly models a process by which agents might decide whether and how to respond to sunspot realizations, using historical experience to determine the information content of such realizations, in the case of the Azariadis model. Because of the formal analogy discussed above, the results are also immediately applicable to the model of Lucas and Stokey as well. It is assumed that agents observe a finite-state Markov process sunspot variable, and entertain the hypothesis that the current sunspot state might be of use in predicting the rate of inflation between the current period and the next. It is also assumed that they use the observed sample distribution of rates of inflation, when the current sunspot state has occurred in the past, as their estimate of the distribution of possible rates of inflation in the current instance. Under such a learning procedure, it is possible for the quantity-theoretic steady state to be unstable, and for stationary sunspot equilibria to be locally stable. In fact, in the case of the policy experiment described above, it is easy to exhibit examples in which, for all values of g above a critical growth rate, the steady state is the unique equilibrium and is stable under the learning dynamics, while for all values of g between β and the critical growth rate, the steady state is unstable and there exist two stationary sunspot equilibria, both of which are locally stable. Hence learning dynamics of this sort are consistent with the conclusion above that reducing g can destabilize the steady-state equilibrium so as to generate fluctuations that reduce expected utility to a level below that associated with the higher rate of money growth.

This analysis shows that agents need not begin with an expectation of exactly the pattern of response associated with a sunspot equilibrium in order for such an equilibrium to come about, assuming that all agents happen to be paying attention to the same sunspot variable. It does not, however, solve the problem of the large number of sunspot variables that agents might equally well pay attention to. In fact, in order for the quantity-theoretic steady state to be unstable, it is necessary that a sufficiently large fraction of the population all be using the same sunspot variable to forecast with; there must be something especially salient about that particular variable, in order to justify a large enough number of agents' paying attention to it for the positive feedback to be strong enough to create self-justifying fluctuations.

Perhaps the most likely case is that in which the variable in question is not a *pure* sunspot variable at all, but rather a very small shock to fundamentals. There may be multiple rational expectations equilibria, in all of which agents respond only to this real shock; and some may involve a response quite out of proportion to the magnitude of the real shock. Equilibria of the latter sort are examples of instability due to self-fulfilling expectations, as much as are sunspot equilibria proper; indeed, sunspot equilibria may be usefully viewed as simply a limiting (and especially dramatic) case of "over-response" of this sort.

The conditions under which stationary sunspot equilibria exist in the Azariadis model or the Lucas-Stokey model immediately translate (via a continuity argument) into conditions under which there exist equilibria exhibiting over-response to real shocks. Furthermore, the results concerning stability of learning dynamics immediately translate as well into conditions under which the equilibrium with a small response to the shocks is unstable and equilibria involving over-response are locally stable. In such a case there is no mystery about why agents should all come to use that particular state variable in forecasting; agents must respond to the variable in *any* rational expectations equilibrium. Then the finding that learning dynamics can lead agents away from the equilibrium in which small shocks to fundamentals have only a small effect indicates that self-fulfilling revisions of expectations may be a realistic source of economic instability.

REFERENCES

Azariadis, Costas, "Self Fulfilling Prophecies," *Journal of Economic Theory,* December 1981, *25*, 380–96.

Balasko, Yves, "Extrinsic Uncertainty Revisited," *Journal of Economic Theory*, October 1983, *31*, 203–210.

Brock, W. A. and Scheinkman, J. A., "Some Remarks on Monetary Policy in an Overlapping Generations Model," in J. H. Kareken and N. Wallace, eds., *Models of Monetary Economies*, Minneapolis: Federal Reserve Bank, 1980.

Cass, David, "A Simple Overlapping Generations Example with Two Interpretations," unpublished, University of Pennsylvania, October 1981.

Diamond, Douglas W. and Dybvig, Philip H., "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economics*, June 1983, *91*, 401–19.

Lucas, Robert E., Jr. and Stokey, Nancy L., "Money and Interest in a Cash-in-Advance Economy," *Econometrica*, forthcoming 1987.

Shell, Karl, "Monnaie et Allocation Intertemporelle," C.N.R.S. seminaire de E. Malinvaud, Paris, November 1977.

Smith, Bruce D., "Legal Restrictions, 'Sunspots', and Peel's Bank Act," unpublished, University of California-Santa Barbara, October 1986.

Woodford, Michael, (1986a) "Stationary Sunspot Equilibria in a Finance Constrained Economy," *Journal of Economic Theory*, October 1986, *40*, 128–37.

______, (1986b) "Learning to Believe in Sunspots," C. V. Starr Center Working Paper No. 86–16, New York University, June 1986.

______, (1986c) "Sunspot Equilibria and Stabilization in a Cash-in-Advance Economy," unpublished, University of Chicago, December 1986.

______, "Expectations, Finance, and Aggregate Instability," in M. Kohn and S. C. Tsiang, eds., *Finance Constraints, Expectations, and Macroeconomics*, New York: Oxford University Press, forthcoming 1987.

DETERRENCE AND ENFORCEMENT OF LAWS[†]

On the Issue of Causality in the Economic Model of Crime and Law Enforcement: Some Theoretical Considerations and Experimental Evidence

By ISAAC EHRLICH AND GEORGE D. BROWER*

The identification of the true causal relationship underlying an apparent empirical association between specific variables of interest is a central issue of concern in all statistical analyses, but especially in social science research where the opportunity to conduct controlled experiments is not available as a rule. In econometric investigations of the incidence of criminal activities the problem concerns, in part, the identification of "supply-of-offenses" and "demand-for-enforcement or protection" functions from scatters of observations concerning crime rates and deterrence variables. While an increase in severity and certainty of punishment is expected to cause a reduction in offenses, an autonomous rise in the risk of victimization is expected to increase public and private enforcement and protection activities, and thus the magnitude of deterrence variables, and generally lower the prospects of illegitimate payoffs.[1]

Another type of simultaneous equation bias in the estimation of deterrent effects, one due to "crowding effects," has also been recognized from the outset (see Ehrlich, 1974). Specifically, if the volume of offenses increases due to changes in exogenous factors, then the extra load on law enforcement agencies could decrease their effectiveness, and thereby cause a reduction in arrest and conviction risks and related sanctions. Empirical observations regarding the actual association between crime and law enforcement variables thus lend themselves in principle to three possible causal interpretations: 1) "punishment (in the sense of all deterrence variables) deters (prevents) crime," as would be revealed by the proper identification of supply of offenses functions; 2) "punishment (in the sense of all disincentives to criminal activity) is raised to meet the threat of crime," as would be revealed by the proper identification of a public response, or a private and public demand for protection, function; 3) "crime deters punishment," as would be revealed by the identification of production functions of law enforcement activities. Some critics of the research into the deterrence hypothesis made this third causal relationship the focal point of their reservations about empirical estimates of the deterrent effect of law enforcement activity (see, in particular, A. Blumstein et al., 1978, and S. A. Hoenack and W. C. Weiler, 1980). They have suggested that the thrust of the econometric findings concerning deterrence may reflect by and large the deterrent effect of crime on punishment rather than any of the other causal relationships just discussed.

The problem of confounding causal relationships can be addressed systematically by properly separating endogenous and exogenous or predetermined variables, imposing

[†] *Discussants*: Michael K. Block, U.S. Sentencing Commission; Philip J. Cook, Duke University; David D. Friedman, Tulane University.

*Melvin H. Baker Professor of American Enterprise and Professor of Economics, SUNY-Buffalo, Buffalo, NY 14260, and Assistant Professor of Business Environment and Policy, University of Connecticut, Storrs, CT 06268, respectively. Ehrlich acknowledges support by grant no. 83-BJ-CX-0013 from the Bureau of Justice Statistics (BJS). He acknowledges the valuable assistance and support by Dr. Steven Schlesinger, Director of BJS. We both acknowledge the useful comments by the discussants in our session.

[1] For some recent expositions of a comprehensive "market model" of illegitimate activities, see Ehrlich (1981).

on the econometric model "identification restrictions," and employing relevant simultaneous equation estimation techniques. Several empirical implementations of the economic model of crime have employed such techniques and obtained similar findings in support of the deterrence hypothesis. Critics have argued that some of the identification restrictions have been incomplete or questionable, but they have not always tested their own arguments. Recently, a number of authors have conducted direct tests of overidentification restrictions and related tests of technical identification mainly in reference to Ehrlich's 1975 study of homicide in the United States. A short overview of these studies and a new attempt at testing the identification restrictions used in Ehrlich's study is contained in Section I. It appears that when a proper test methodology is used, one cannot reject the hypothesis that the latter restrictions are valid, and this conclusion is strengthened by inference from alternative tests of Granger causality conducted by L. Phillips and S. C. Ray (1982).

While the reported tests appear to be consistent with basic theoretical expectations, we consider them but a first step toward a satisfactory analysis of the issue of causality in crime statistics, since they address only the technical (order and rank) conditions for overidentification restrictions in a *limited* econometric setting. What is desirable for a fuller, or *substantive* identification, is the identification and measurement of key exogenous factors that are responsible for significant "autonomous" shifts in demand or production schedules which, in turn, can trace the supply-of-offenses function believed to be identified.

The issue can be presented more sharply in the context of a time-series analysis of all crime variations in the United States. If the specific variations in the clearance and conviction rates (or in the execution risk for the crime of murder) are found to be largely autonomous, what are the unique exogenous factors that influence these variations over time?

A possible answer may be obtained by investigating more substantively than heretofore attempted in the literature the nature of the "production" of law enforcement by police and courts. This issue is briefly explored in Sections II and III. We hypothesize that a major "efficiency parameter" which affects the production of police and court activity, and thus the conditional probabilities of apprehension, conviction, and punishment, is the *legal environment* prescribing the activities of the corresponding law enforcement agencies: the rules of evidence, due process, and defendants' rights as determined primarily by the Supreme Court of the United States. The hypothesis is that the Court, through its creation or certification of legal precedent, or "legal capital," may exert critical influence on the magnitude of variations in deterrence variables at any given levels of public appropriations to law enforcement activity and offense rates. The focus of our analysis in this part is twofold. In Section II we first consider the attempt to quantify Court rulings that may serve as an overall index of the changing "due process environment" in a direction of either greater ease or difficulty of achieving apprehension and conviction of felons under a given budget and efficiency of law enforcement resources. We then turn in Section III to the estimation of production functions of police and court activity that include the latter index.

I. Technical Identification: Testing Overidentification Restrictions in the Presence of Serially Correlated Errors

Hoenack and Weiler (henceforth H-W), argued that Ehrlich's (1975) supply of murders function was not identified. They reached this conclusion by implementing an F-test for overidentification restrictions proposed by R. L. Basmann (1960). However, their test statistic is improper since it does not account for the serial correlation in error terms confirmed in both Ehrlich's (1975) and Stephen Layson's (1985) studies. Moreover, it is based on a semilogarithmic specification, which has been found an inefficient functional specification in both Ehrlich's (1977) and Layson's studies. Implementing a Box-Cox test procedure for optimal functional form, both studies have confirmed the superiority of a log-log transformation as an optimal regression specification. Furthermore, H-W's rejection of the deterrence hypothesis is based in part on an *ad hoc* inclusion of multiple age composition variables which are highly correlated, are not found to be jointly significant as a rule, and in any event have little bearing on the identification issue they raised. These shortcomings are also noted by Layson, but he does not provide an alternative test statistic for overidentification restrictions.

The supply of murder equation has been specified in Ehrlich (1975) as follows:

$$(1) \qquad y_1 = Y_1 b_1 + X_1 c_1 + v_1,$$

where

$$(2) \qquad v_1 = r v_{1-1} + e_1.$$

In equation (1), y_1 denotes the natural logarithm of the offense rate (q); Y_1 the natural logarithms of the right-hand side endogenous variables entering the murder equation: the conditional probabilities of apprehension (Pa), conviction ($Pc|a$) and execution ($Pe|c$); and X_1 the natural logarithms of the rate of unemployment (U), labor force participation (L), fraction of residential population age 14–24 (A), and permanent income (Y_p), as well as the chronological time T (1933 = 1 through 1967 = 37), and a constant term (C).

This equation has been identified through a vector of exogenous variables, X_2, including the total civilian population (N), annual per capita (real) domestic expenditures of all governments ($XGOV$), annual per capita (real) expenditures on police lagged one year ($XPOL_{-1}$), and the fraction of nonwhites in the population (NW). These variables have been excluded from equation (1) on the grounds that they belong in at least one of the other structural equations of the model explaining the "production" of deterrence variables and the allocation of public resources to law enforcement activity.

The a priori justification for these identification restrictions (which satisfy the order conditions) follows from the logic of the economic approach, which focuses on the role of incentives in crime and some previous experience with the empirical implementation of this approach, especially in so far as demographic variables are concerned. For example, $XPOL$ is a measure of the level of resource inputs devoted to law enforcement activities. It does not represent a direct "incentive" (price, tax, or reward) which the theory identifies as having direct bearing on the decision to participate in criminal activity. While both age and racial compositions of the population may account for differential opportunities or propensities to participate in crime, at least in this sample NW was found to have no impact on q when introduced jointly with T, although it was found to be systematically related to the production of deterrence variables.

The validity of these overidentification restrictions can be tested using a modified Basmann test statistic. Given the presence of serial correlation, equation (1) can be rewritten as

$$(3) \quad (y_1 - r y_{1-1}) = (Y_1 - r Y_{1-1}) b_1 + (X_1 - r X_{1-1}) c_1 + e_1,$$

$$\text{or} \qquad y_1^* = Y_1^* b_1 + X_1^* c_1 + e_1,$$

and the identification restrictions can be addressed by testing whether the set of omitted variables X_2 does not in fact belong in equation (1), that is, by testing the null and the alternative hypotheses defined as H_0: $d_1 = 0$ and H_1: $d_1 \neq 0$ in the equation

$$(4) \quad y_1^* = Y_1^* b_1 + X_1^* c_1 + X_2^* d_1 + e_1,$$

$$\text{where} \qquad X_2^* = X_2 - r X_{2-1}.$$

Since in equations (3) and (4) e_1 has all the desired normal properties, the Basmann statistic would apply to alternative estimates of var(e_1) in the context of the modified first-difference transformations of the structural and reduced-form equations, with the coefficient of serial correlation estimated through an iterative search procedure. Applying Basmann's reasoning,

$$(5) \quad [(\hat{e}_1'\hat{e}_1 - \hat{a}_1'W^*\hat{a}_1)/\hat{a}_1'W^*\hat{a}_1] \cdot [(T - K)/(K - N_1)]$$

has an approximate F-distribution with $(K-N_1)$ and $(T-K)$ degrees of freedom, where T denotes the total number of observations, K the number of all exogenous variables in the structure, and N_1 the total number of right-hand side variables in equation (1). In equation (5) the first term in the numerator of the first ratio could be estimated jointly with $\hat{r}$ and the coefficient vector $\hat{b}_1$ applying a conventional two-stage least square ($TSLS$) procedure to equation (3) for alternative values of $-1 \leq \hat{r} \leq 1$. The search procedure would result in the selection of $\hat{r}$, $\hat{b}_1$, and $\hat{c}_1$ that minimized the sum of squared residuals in that equation. The second term in that numerator, and hence the denominator, would then be computed using the identity $\hat{a}_1 = [1, -\hat{b}_1]$ and the sum of squared residuals in the modified reduced-form equations $W^* = \hat{V}^{*\prime}\hat{V}^*$, where $\hat{V}^* = [Y^* - X^*(X^{*\prime}X^*)^{-1}X^{*\prime}]$. We shall refer to the resulting statistic in equation (5) as the modified Basmann (MB) statistic.

An alternative measure of the desired statistic summarized in equation (5) can be achieved by calculating the values of $\hat{r}$ and $\hat{e}_1'\hat{e}_1$ using the three-round estimation procedure proposed in Ray Fair (1970), which employs an expanded reduced-form regression in the first stage of the analysis. The second term in the numerator of (5), and thus the denominator of that equation, would then be calculated using Fair's estimates of $\hat{a}_1$ and the term $W^* = \hat{V}^{*\prime}\hat{V}^*$, with the reduced-form regressions modified by Fair's estimate of $\hat{r}$. We shall refer to the resulting estimate as Fair-modified Basmann (FB) statistic. Both measures permit consistent tests of the validity of the overidentification restrictions in the presence of serially correlated residuals.

TABLE 1—SUPPLY OF MURDERS EQUATION: EHRLICH'S (1975) DATA[a]

Algorithms used for Stat, $\hat{b}$	Pe/c Estimate			
	PXQ_1		PXQ_2	
	FB, Fair (1)	MB, $TSLS$ (2)	FB, Fair (3)	MB, $TSLS$ (4)
Pa	−1.284	−2.116	−1.089	−1.765
	(−1.804)	(−2.109)	(−1.170)	(−1.991)
$Pc\|a$	−0.4358	−0.3263	−0.3852	−0.2667
	(−3.638)	(−1.759)	(−3.835)	(−1.693)
$Pe\|c$	−0.0268	−0.0364	−0.0628	−0.0642
	(−1.148)	(−1.472)	(−3.314)	(−3.321)
Y_p	1.534	1.360	1.368	1.174
	(4.041)	(3.536)	(4.712)	(3.731)
L	−1.767	−0.8969	−1.471	−0.0828
	(−1.999)	(−0.8222)	(−1.882)	(−0.8910)
U	0.0613	0.0669	0.0612	0.0608
	(1.833)	(2.041)	(2.246)	(2.274)
A	0.7501	0.5569	0.5378	0.4077
	(2.489)	(1.826)	(2.347)	(1.726)
T	−0.0447	−0.0353	−0.0468	−0.376
	(−4.001)	(−4.022)	(−6.052)	(−5.360)
C	−4.887	−0.1954	−4.966	−0.8440
	(−1.334)	(−0.0456)	(−1.522)	(−0.2090)
Rho	0.4672	0.1997	0.2648	0.1347
SSR	0.0608	0.0690	0.0454	0.0482
SER	0.0484	0.0595	0.0418	0.0431
Stat	1.139	2.234	2.322	2.512
d.f.	2,24	2,24	2,24	2,24

[a] Dependent Variable = q (Murder rate); $\hat{b}/S\hat{b}$ in parentheses. The reported coefficient estimates and (implied) standard errors differ slightly from those of Ehrlich (1975, Table 3) because the computer program used was modified. The Beach and Mackinnon algorithm has replaced the Cochrane and Orcutt procedure for regression with serially correlated errors due to the former's greater efficiency.

The set of reported tests refers to specific regressions conducted in Ehrlich's (1975) study. Estimated supply-of-murder functions are illustrated in Table 1 for the variables defined in equation (1) above and for two alternative measures of $Pe|c$: (1) $PXQ_1 = E_{t+1}/C_t$; (2) $PXQ_2 = E_t/C_t$; where E_{t+1} and C_t, for example, denote the number of executions and convictions in years $t+1$ and t, respectively. The deterrence hypothesis is supported by all regressions. The values of the test statistics MB and FB, shown in the row named "Stat," do not reject the null hypothesis that the coefficients of the variables excluded from the murder-supply function (2) are zero, since the F values are always below the critical levels of F at the 5 percent significance level.

The main limitation of these results, however, stems from the paucity of exogenous variables that could adequately account for autonomous shifts in the key deterrence variables represented by Y_1 in equation (1). In the next sections we experiment with the construction and use of an index of the legal environment underlying law enforcement activity as such an exogenous variable.

II. Modeling the Role of the Legal Environment in Influencing Law Enforcement Outcomes

The constitutional autonomy of the U.S. Supreme Court makes it plausible to assume that the actions of the Court are taken largely independently from contemporaneous variations in crime rates or deterrence variables. Although the Court may be influenced by social, economic, and scientific or cultural trends, there is little reason to believe that in its constitutional charge it aims to serve as a proxy for the common will of the moment, let alone year-to-year changes in specific crimes and law enforcement activities. Indeed, in the context of understanding the behavior of the criminal justice system, the role of the Court may resemble that of the Federal Reserve Board as a relatively autonomous institution.

Isolating the influence of the high court from that of other state courts and quantifying the consequences of its decisions on the legal environment bearing on due process is subject to myriad limitations. Among these limitations are questions such as whether Court rulings lead, simply ratify, or lag rulings of state courts, and whether its decisions have constant, augmenting, or decaying residual influence on future decisions by the Court itself or lower courts. The quantitative indexes that are more simple to construct are also unidimensional: they are not weighted separately for different aspects of due process. We believe, however, that all such difficulties should not preclude modest attempts to quantify and test the influence of those Court decisions which are believed, on the whole, to have either expanded or limited the scope of civil liberties for offenders relative to the status quo ante. In pursuing such work we have surveyed the opinions of legal scholars and experts in criminal law as to which Court decisions should be included, and how to weight their significance as changing the status quo in the direction of expanding or limiting offenders' rights in all aspects of due process.

TABLE 2—SIGNIFICANT COURT DECISIONS, 1933–77

Decision	Date
Powell v. Alabama, 287 U.S. 45	11–07–32
Brown v. Mississippi, 297 U.S. 587	2–17–36
Chambers v. Florida, 309 U.S. 227	2–12–40
Waley v. Johnston, 316 U.S. 101	4–27–42
Brown v. Allen, 344 U.S. 443	4–09–53
Griffin v. Illinois, 351 U.S. 12	4–23–56
Mallory v. U.S., 354 U.S. 449	6–24–57
Spano v. New York, 360 U.S. 315	6–22–59
Mapp v. Ohio, 367 U.S. 643	6–19–61
Townsend v. Sain, 372 U.S. 293	3–18–63
Gideon v. Wainwright, 372 U.S. 335	3–18–63
Douglas v. California, 372 U.S. 353	3–18–63
Fay v. Noia, 372 U.S. 391	3–18–63
Malloy v. Hogan, 378 U.S. 1	6–15–64
Escobedo v. Illinois, 378 U.S. 478	6–22–64
Miranda v. Arizona, 384 U.S. 436	6–13–66
U.S. v. Wade, 388 U.S. 218	6–12–67
Katz v. U.S., 389 U.S. 347	12–18–67
Furman v. Georgia, 408 U.S. 238[a]	6–29–72
Davis v. U.S., 411 U.S. 233	2–20–73
Francis v. Henderson, 425 U.S. 537	5–03–76
Doyle v. Ohio, 426 U.S. 610	6–17–76
Gregg v. Georgia, 428 U.S. 153[a]	7–02–76
Stone v. Powell, 428 U.S. 465	7–06–76
Wainright v. Sikes, 433 U.S. 72	8–29–79

[a] Since these cases carry significance only for murder, they have not been included in the aggregate index, but were entered (cumulatively) in the general index constructed specifically for murder.

The actual construction of alternative indexes was made in two stages. Based on official government publications, Brower collaborating with Jacqueline Lauriah and advisers at the School of Law at Buffalo (SUNY) constructed a list of important Court cases and weighted it according to standard decision analysis procedures. Alternative indexes were later constructed from questionnaires distributed among leading legal scholars at other universities. They were asked to add or delete cases from the original list and to weigh each case on a scale of the integers 0 to +3 or −3, depending upon whether cases were considered as "unimportant," "significant," "important," or "land-mark" in the direction of bolstering or limiting offenders' rights. The constructed index applying to all crime categories in each year was obtained by *summing* the annual weights assigned to the cases which have been decided between 1933 and that year, and a correction was made for the portion of the year the (dated) ruling has been in effect.[2] That is, the index is a measure of the stock of legal capital expanding the scope of

[2] Of the 27 questionnaires sent, we received 6 responses.

suspected offenders' rights, under the assumption of zero rate of decay of rules which remain in effect. Table 2 includes the expanded list of cases identified by all respondents.

We emphasize that our construction of an experimental index is an exercise in "positive methodology." It is void of any normative implications regarding the wisdom and desirability of the decision. Indeed, Court decisions which bolster individual rights may reduce the risk of "legal error" and thus improve justice as well as deterrence. However, significant changes in the status quo ante raise the costs of all apprehensions and convictions, and may thus be expected to lower the productivity of law enforcement efforts (at given budgets) as well as the optimal magnitudes of relevant enforcement variables, as sometimes noticed by the Court itself.[3]

III. Production Function Estimates: A Summary of Results

The police and court production functions are estimated as part of a simultaneous equation model of crime and law enforcement, as implemented in Ehrlich (1974, 1975) and Brower. The variable definitions and basic structure are those used in Ehrlich (1975).[4] We now estimate, however, both police production functions for probabilities of apprehension for specific crimes (Pa_i), using data from 1933 to 1977, and court production functions for the conditional probabilities of conviction given charge, ($Pc|a_i$), using observations from 1946 to 1977, since data on real per capita court expenditures by all governments ($XCOURT$) are available only for that period. Offense and deterrence variables data are expanded to include all reported crime categories. The quality of the offense data for specific crimes (especially rape, larceny and auto theft) and the conviction data for all is highly questionable, despite the availability of the revised FBI offense data for 1933–72. However, extensive tests of stationarity and stability of the time-series on offenses and deterrence variables conducted by Brower indicate no apparent technical reasons to preclude their employment a priori. The time-series on police expenditures, implicit price deflators, and permanent income are also updated and revised according to the latest Department of Commerce revisions.

Each production function (y) is specified in log-log form. It includes the relevant resource variables: police employees per capita ($MPOL$) for Pa, and real per capita expenditures on police and courts ($XPOL$ and $XCOURT$) for $Pc|a$; the relevant loading factors: offense rate (q) for Pa and arrest rate (Arr) for $Pc|a$; demographic variables (N, A, NW) and the time trend (T). A consensus Court index (CI)[5] is introduced as an efficiency parameter in both, although an alternative dummy Court index (DU Index), separating the intensive reform period of 1961–63 (see Table 2) from other years is also used in the court production function. In some aggregate court production regressions $Pc|a$ is replaced by a measure of the probability of going to prison given conviction ($Pj|c$) or the expected prison term conditional on conviction ($Ej|c$), estimated from data on total admissions and year-end populations in state and local prisons.[6]

Selected estimates of the production functions are shown in Tables 3 and 4. Police production estimates generally have the ex-

[3] See, for example, Justice Powell's comment at 408 U.S. 238, 435 n. 18 (1972) on the effects of *Mapp v. Ohio* and *Miranda v. Arizona* and related federal habeas corpus decisions on the frequency of executions in the 1960's.

[4] Due to space limitations we relegate an elaborate description of variables, source data, and the comprehensive structure and overidentification restrictions to an appendix which is available to the reader upon request.

[5] The consensus index (CI) is computed by averaging the annual Court index weights that are derived from the independent scaling of cases by individual respondents (including Brower and Lauriah).

[6] The resource variables and arrest and crime rates are treated as endogenous variables. Excluded exogenous variables are $XPOL_{-1}$ (or $MPOL_{-1}$), $XCOURT_{-1}$, and Y_p, L, U, and $XGOV$ as defined in Section I. Regression estimates are derived via Fair's three-round estimation procedure.

TABLE 3—POLICE PRODUCTION FUNCTIONS SAMPLE RESULTS[a]

	Offense Category			
	Murder	Robbery	Robbery	Burglary
Sample Period	1935–77	1935–77	1946–77	1946–77
q	−0.131	−0.156	−0.167	−0.022
	(−2.920)	(−2.690)	(−1.690)	(−0.119)
MPOL	0.149	0.278	0.646	0.192
	(1.460)	(1.270)	(2.690)	(0.950)
Index	0.001	−0.015	−0.012	−0.016
	(0.578)	(−2.880)	(−2.090)	(−2.870)
Trend	0.003	−0.028	−0.121	−0.119
	(0.817)	(−3.610)	(−2.080)	(−2.790)
N	0.304	2.240	5.740	6.350
	(1.530)	(5.230)	(2.490)	(3.930)
A	−0.045	0.097	0.167	−0.007
	(−0.459)	(0.475)	(0.678)	(0.025)
NW	−1.600	−0.544	3.870	2.290
	(−4.100)	(−0.592)	(1.360)	(1.111)
C	−0.411	33.30	162.0	155.0
	(−0.082)	(2.510)	(2.010)	(2.540)
Rho	0.091	−0.066	−0.094	0.237
SSR	0.012	0.069	0.054	0.035
SER	0.018	0.044	0.047	0.038

[a] $\hat{b}/S\hat{b}$ in parentheses.

TABLE 4—COURT PRODUCTION FUNCTIONS, 1946–77[a]

	y and Index			
	P(*C*|*A*), *DU*	Index	*P*(*J*|*C*), *CI*	*E*(*J*|*C*), *CI*
Offense Category	Total	Persons	Total	Total
Arr	0.450	−0.255	−1.310	−1.270
	(1.690)	(−0.307)	(−2.820)	(−2.760)
XCOURT	0.184	−0.784	−0.524	−0.009
	(1.500)	(−1.050)	(−0.497)	(−0.009)
XPOL	0.250	−0.139	0.145	0.142
	(1.500)	(−0.511)	(0.539)	(0.488)
Index	−0.111	−0.218	−0.047	−0.036
	(−2.250)	(−2.760)	−2.550)	(−1.970)
Trend	0.132	0.033	−0.087	−0.102
	(2.350)	(0.306)	(−0.908)	(−1.060)
N	−11.700	−2.950	9.230	9.600
	(−4.470)	(−0.746)	(1.850)	(2.000)
A	−2.360	1.980	3.100	2.070
	(−3.580)	(1.980)	(2.200)	(1.470)
NW	−0.670	3.590	3.440	−0.199
	(−0.184)	(0.393)	(0.553)	(−0.031)
C	−109.00	−46.600	41.700	79.400
	(−1.460)	(−0.317)	(0.325)	(0.607)
Rho	0.361	0.250	0.574	0.439
SSR	0.077	0.163	0.230	0.240
SER	0.058	0.035	0.100	0.102

[a] See Table 3.

pected signs for the resource and load variables and are often significant. The Court index coefficient also has the expected negative sign and is generally significant. Court production functions for groups of offenses exhibit similar albeit weaker results. The results for the resource variables are generally insignificant when the dummy Court index is employed, but the latter index's effect is generally significant. In addition, the consensus Court index has the expected sign and significance when the dependent variable is $Pj|c$ or $Ej|c$. We believe that the generally weak results obtained for the Court production function for $Pc|a$ using the consensus index (not shown here) are due to the special role of bargaining and pretrial agreements in determining the probability of conviction given charge for crime. Many due process reforms may be internalized rather rapidly by prosecutors and attorneys through agreements to plead to a lesser crime or receive a shorter sentence upon conviction. Indeed, $Pc|a_i$ may even increase as a result of specific Court rulings which expand the scope of offenders' rights, since once a charge or expected sentence is bargained, the odds of obtaining conviction may rise and the expected sentence may fall. This may be the reason why the dummy Court index "works": the very intensive reforms in due process in 1961–63 may have generated a once and for all reduction in $Pc|a_i$. This may also be the reason why the consensus Court index (CI) has a significant effect on $Pj|c$ and $Ej|c$.

Finally, we have also tested the influence of the (lagged) Court index, in both its general form and the special one constructed for murder, on the conditional probability of execution given conviction, using a simple ordinary least square regression routine and a correction for serial correlation with and without a trend variable. The coefficients are around −0.2 and the T-ratios are 10 or above. Apparently, as suggested by Judge Powell (see fn. 2), Court decisions have been mostly responsible for the movements in execution risks overtime.

This analysis must be regarded as highly tentative in view of the myriad limitations inherent in assembling and interpreting the legal data and their relevance for different offenses and various stages of the enforcement process, and because of the partial nature of the econometric model, severe limitations of the data, the collinearity among explanatory variables, and the shortness of the time-series. The results obtained so far indicate only the potential usefulness of studying the role of the legal environment in explaining autonomous movements in basic deterrence variables and thus possibly in the corresponding trends of crime as well.

REFERENCES

Basmann, R. L., "On The Finite Sample Distribution of Generalized Classical Linear Identifiability Test Statistics," *Journal of the American Statistical Association*, December 1960, *55*, 650–59.

Blumstein, A. et al., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, Washington: NAS, 1978.

Brower, G. D., "The Supreme Court and the Growth of Crime," unpublished doctoral dissertation, SUNY-Buffalo, 1985.

Ehrlich, Isaac, "Participation in Illegitimate Activities—An Economic Analysis," in G. S. Becker and W. M. Landes, eds., *The Economics of Crime and Punishment*, New York: Columbia University Press, 1974, 68–134.

______, "The Deterrent Effect of Capital Punishment—A Question of Life and Death," *American Economic Review*, June 1975, *65*, 397–417.

______, "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," *Journal of Political Economy*, August 1977, *85*, 741–88.

______, "On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence," *American Economic Review, June* 1981, *71*, 307–32.

Fair, R. C., "The Estimation of Simultaneous Equation Models with Lagged Endogeneous Variables and First Order Serially Correlated Errors," *Econometrica*, May 1970, *38*, 507–16.

Hoenack, S. A. and Weiler, W. C., "A Structural Model of Murder Behavior and the Criminal Justice System," *American Economic Review*, June 1980, *70*, 327–41.

Layson, Stephen K., "Homicide and Deterrence: A Reexamination of the United States Time-Series Evidence," *Southern Economic Journal*, July 1985, *52*, 68–89.

Philips, L. and Ray, S. C., "Evidence on the Identification and Causality Dispute About the Death Penalty," in O. D. Anderson and M. R. Perryman, eds., *Applied Time Series Analysis*, New York: Elsevier, 1982, 313–40.

A Model of Optimal Incapacitation

By STEVEN SHAVELL*

One of the functions of the criminal sanctions of imprisonment and the death sentence is to prevent individuals from doing harm by removing them from the population.[1] This incapacitative function of sanctions is considered below in a model in which the amount of harm individuals cause each period that they are free is not influenced by the threat of sanctions (so as to abstract from the role of sanctions as a deterrent).[2]

The model is initially examined assuming that an individual's dangerousness (i.e., the harm he will do if free) remains the same each period of his life, and that the sanction is imprisonment. In this case, it is optimal to imprison an individual if his dangerousness exceeds a threshold equal to the per period social cost of imprisonment. Moreover, if it is optimal to imprison an individual at all, it will be best to do so for life. The optimal probability of apprehension is also determined.

The model is then extended in several ways. It is first supposed that the dangerousness of individuals declines with age. In this case, it is again optimal to imprison individuals if their dangerousness exceeds the per period cost of imprisonment, but it is optimal to release them if their dangerousness later falls below the threshold. It is next supposed that the dangerousness of individuals declines with time spent in prison due to a rehabilitative effect. In this case, it is optimal to imprison individuals beginning at a lower threshold of dangerousness than the per period cost of imprisonment (imprisonment is now socially more valuable) and to release them if their dangerousness becomes sufficiently low. Finally, it is supposed that a choice can be made between imprisonment and the death penalty, and the optimal choice is discussed.

I. Analysis

A. *Basic Model*

An equal number (normalized at one) of finite-lived individuals enter the population every period, so that in the steady state the population is comprised of equal numbers of individuals of each age cohort. An individual causes the same amount of harm every period that he is free, where the amount of harm varies by the individual. Individuals are apprehended each period with a probability, and if apprehended, may be imprisoned. Specifically, let h = harm done each period by an individual if not imprisoned; $0 \leqq h \leqq \bar{h}$; $f(h)$ = probability density of individuals of type h entering the population each period; n = number of periods that each individual lives; p = probability of apprehension of individuals each period; $s(h)$ = length of prison sentence imposed on an individual of type h if apprehended.

Society bears certain costs in apprehending and imprisoning individuals: let c = cost of imprisoning an individual per period,[3] $c > 0$; $e(p)$ = enforcement expenses associated with maintaining the probability of apprehension; $e'(p) > 0$, $e''(p) > 0$.

*Professor of Law and Economics, Harvard Law School, Cambridge, MA 02138. I thank Michael Block, Philip Cook, David Friedman, Louis Kaplow, and A. Mitchell Polinsky for comments and the National Science Foundation (grant no. SES-8420226) for financial support.

[1] Note that monetary sanctions have no such function.

[2] While the deterrent effect of sanctions has of course been much investigated by economists (see fn. 5), the incapacitative role of sanctions does not appear to have been the focus of a theoretical study. However, Isaac Ehrlich (1981) contains an interesting section on incapacitation which emphasizes so-called replacement effects; there is little overlap between his paper and this one.

[3] This may be interpreted as including not only the costs associated with the building and operation of prisons, but also the forgone production of the imprisoned individual and the disutility he suffers.

The social problem is to choose a system of sentences and the probability of apprehension to minimize expected social costs, defined as the expected sum of harm, the costs of imprisonment, and enforcement expenses.

The optimal system of sentences is clear: an apprehended individual for whom $h > c$ will be imprisoned for life, but one for whom $h \leqq c$ will be set free. This is because individuals for whom $h > c$ do more harm any period that they are free than they cost to imprison, and conversely if $h \leqq c$. (If $h = c$, it does not matter whether individuals go free, but for simplicity I adopt the convention that they do.)

Note that the optimal sentences do not depend on p or on n, but they do depend on c.

Given that sentences are optimal, social costs per period as a function of the probability of apprehension are

$$(1)\quad n\int_0^c hf(h)\,dh + q(p)\int_c^{\bar{h}} hf(h)\,dh + (n - q(p))\int_c^{\bar{h}} cf(h)\,dh + e(p),$$

where

$$(2)\quad q(p) = (1-p) + (1-p)^2 + \ldots + (1-p)^n.$$

That is, the sum of the probabilities that individuals in different age cohorts have not been apprehended, and thus where the sum of the probabilities of individuals in the different cohorts who have been apprehended is

$$(3)\quad [1-(1-p)] + \ldots + [1-(1-p)^n] = n - q(p).$$

The first-order condition determining the optimal p is therefore

$$(4)\quad e'(p) = -q'(p)\int_c^{\bar{h}}(h-c)f(h)\,dh.$$

Namely, marginal enforcement expenses must equal the reduction in harm (net of the cost of imprisonment) due to imprisonment of additional individuals.

It follows from (4) that the optimal probability rises with n; since $-q'(p) = 1 + 2(1-p) + \ldots + n(1-p)^{n-1}$ rises with n, p must rise to maintain equality in (4). (If individuals do harm for a longer time, it is more important to incapacitate them.) Also the optimal probability rises with a rightward shift in the distribution of h; if the harm done by a person of type h is $h + t$, where t is a parameter, then the right-hand side of (4) becomes $-q'(p)\int_{c-t}^{\bar{h}}(h+t-c)f(h)\,dh$. Differentiating this with respect to t gives $-q'(p)\times$(Probability that h exceeds $c-t$), which is positive. Hence, again, p must rise to maintain equality in (4). In addition the optimal probability rises with a decline in c; differentiating the right-hand side of (4) with respect to c gives $q'(p)\times$ (Probability that h exceeds c), which is negative. (Since a decline in c means it is optimal to imprison a greater percentage of individuals who are apprehended, the social payoff from raising the probability is enhanced.)

B. *Harmfulness a Function of Age*

Suppose now that the harm an individual will do if free declines with his age. (There is strong evidence that dangerousness does decline with age; see, for instance, U.S. Department of Justice.) Then it is clear that it will be optimal to imprison an apprehended individual for whom $h > c$ and to release him as soon as $h \leqq c$. The optimal probability of apprehension will be determined by a condition similar to (4) (and will fall) but for simplicity the probability will not be reconsidered here or below.

C. *Rehabilitative Effect of Imprisonment*

Suppose next that time spent in prison will reduce an individual's harmfulness (and, to isolate this rehabilitative effect, suppose that age itself will not reduce harmfulness). In particular, assume that each period an individual is in prison, his harmfulness will be multiplied by a factor r, where $1 > r > 0$; hence if he is in prison i periods and released, the harm he will do each subsequent period will be $r^i h$. It will be shown that the optimal sentencing policy depends on a comparison with the threshold

$$(5) \qquad t(j) = c/[j-(j-1)r],$$

where j is the number of periods remaining in an individual's life. Thus $t(1) = c > t(2) = c/(2-r) > t(3) = c/(3-2r)\dots$. The optimal sentencing policy is to imprison an apprehended individual with j periods left in his life if $h > t(j)$, and to release him if, as may happen, his harmfulness later falls to or below the then-relevant threshold.

Notice that since $t(j) < c$ for $j > 1$, it may be optimal to imprison an individual even though his current harmfulness h is less than the per period cost of imprisonment c. The reason is that the rehabilitative effect of imprisonment means that when he is later freed, he will do less harm; thus imprisonment will benefit society in a way in addition to incapacitating the individual. The rehabilitative effect is more important if an individual is young since society will have more time in which to accomplish rehabilitation and more time in which to enjoy its benefits; thus the threshold rises with age. The rehabilitative effect is of no value to society if an individual has only one period left to live, so it makes sense that $t(1) = c$.

That the optimal sentencing policy is as claimed can be proved by induction. For $j = 1$, it is obvious that it is optimal to imprison an individual if $h > c$ and not otherwise. Consider now an apprehended individual with $j > 1$ periods of life remaining for whom $h < t(j-1)$. We know by the inductive hypothesis that whether or not the individual is imprisoned this period, it will be optimal for him to be free for all periods after this period (since $h < t(j-1)$). Hence, if he is allowed to go free this period, social costs associated with him will be jh, and if he is imprisoned this period, the social costs will be $c + (j-1)rh$. Accordingly, as long as $jh \leqq c + rh(j-1)$, it will be optimal to allow him to go free. Solving $jh = c + rh(j-1)$ for h, we obtain the asserted threshold $c/[j-(j-1)r] = t(j)$. (It was assumed here that $h < t(j-1)$; if h is larger, the argument that it continues to be optimal to imprison an individual is tedious and is omitted.)

D. *Death Penalty vs. Imprisonment*

Suppose last that the situation is as in the basic model except that the death penalty is available as an alternative to imprisonment, where the death penalty involves a social cost d.[4] We know that, in the absence of the death penalty, it is optimal to imprison an individual for life if $h > c$ and to allow him to go free otherwise. This means that, in the absence of the death penalty, for an individual who is apprehended and who will live j more periods, social costs will be augmented by jc if $h > c$ and jh otherwise. Hence, it will be optimal to apply the death penalty if and only if $d < \min(jh, jc)$. Thus, for given j, the death penalty will not be employed for any h if $d \geqq jc$; and if $d < jc$, the penalty will be imposed if $h > d/j$ (and individuals will be set free otherwise). The penalty is more likely to be optimal when j is large since then society will save more in imprisonment costs.

II. Concluding Comment: Optimal Sanctions when Incapacitation is the Goal vs. when Deterrence is the Goal

There are several important differences between the optimal magnitude of sanctions when incapacitation is the goal from when deterrence is the goal.[5] First, when deterrence is the goal, the optimal sanction is generally higher the lower is the probability of apprehending an individual. But as seen here, when incapacitation is the goal, the optimal sanction is independent of the probability of apprehension (although the optimal probability depends on the optimal

[4] The cost d could be interpreted as including not only the resource costs of imposing the death penalty, which might be small, but also an amount reflecting social distaste for the penalty or its disadvantageous effects on the public perception of the value of human life. Note also that the sanction of permanent deportation is, for our purposes, qualitatively identical to the death penalty, since it removes an individual from the population and involves (for the deporting country) a once and for all cost.

[5] See Gary Becker's original paper (1968) on the use of sanctions to deter, and see also R. Carr-Hill and Nicholas Stern (1979), A. Mitchell Polinsky and myself (1984), and my 1985 paper.

sanctioning policy). Second, when deterrence is the goal, the optimal sanction generally rises continuously with the magnitude of the harm done, but when incapacitation is the goal, the sanction rises discontinuously, from zero to a positive and fixed level once a threshold level of harm is surpassed. Third, when deterrence is the goal, the optimal magnitude of sanctions depends on the ability to deter, and if this ability is small (as, for instance, with the enraged), a low optimal sanction will be indicated. But a high sanction might be called for to incapacitate.

REFERENCES

Becker, Gary, "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, March/April 1968, *76*, 169–217.

Carr-Hill, R., and Stern, Nicholas, *Crime, The Police, and Criminal Statistics*, London: Academic Press, 1979, 280–309.

Ehrlich, Isaac, "On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence," *American Economic Review*, June 1981, *71*, 307–22.

Polinsky, A. Mitchell and Shavell, Steven, "The Optimal Use of Fines and Imprisonment," *Journal of Public Economics*, June 1984, *24*, 89–99.

Shavell, Steven, "The Optimal Use of Nonmonetary Sanctions as a Deterrent," Law and Economics Discussion Paper No. 10, Harvard Law School, 1985.

U.S. Department of Justice, *Report to the Nation on Crime and Justice: The Data*, NCJ–87068, Washington: USGPO, October 1983, 32–33.

Permanent and Transitory Components in Macroeconomic Fluctuations

By JOHN Y. CAMPBELL AND N. GREGORY MANKIW*

A "stylized fact" is sometimes defined as an empirical claim that is widely believed but the evidence for which is only mixed. Seen in this light, it is perhaps not surprising that one of the principal stylized facts of macroeconomics has undergone a substantial reexamination over the past five years. The purpose of this paper is to review some of the recent developments that have led to the new view of output fluctuations and then to provide some additional evidence.

According to the traditional view, fluctuations in real GNP primarily reflect temporary deviations of production from trend. Only a few years ago, economists as diverse as Olivier Blanchard (1981) and Finn Kydland and Edward Prescott (1980) subscribed to the then-uncontroversial suggestion that the logarithm of quarterly real GNP is well represented as a stationary second-order autoregressive process around a deterministic time trend. Blanchard estimated the following process for detrended log real GNP:

$$(1) \qquad y_t = 1.34 y_{t-1} - 0.42 y_{t-2} + \varepsilon_t .$$

This AR(2) process implies that the effect of a shock to output increases for a few quarters, but the effect dies out soon thereafter. Only 8 percent of a shock to output remains after 20 quarters.[1] These authors, along with many others, viewed this dynamic response of output to an innovation as an important phenomenon to be explained by macroeconomic theory.

The reexamination of this traditional view was motivated in part by developments in the econometrics of nonstationary time-series (for example, David Dickey and Wayne Fuller, 1981). In one of the first applications of the new econometric techniques to standard macroeconomic data, Charles Nelson and Charles Plosser (1982) argued that most of these time-series, including real GNP, are not stationary around a deterministic trend. Instead, these series should be considered stationary only after differencing. Nelson and Plosser suggested that the data for output are well approximated by simple stochastic processes that imply no trend reversion at all.

Much disagreement remains over exactly how persistent are shocks to output. Nonetheless, among investigators using postwar quarterly data, there is almost unanimity that there is a substantial permanent effect. In a previous paper (1987), we estimated unconstrained ARMA models for GNP growth. We found that a negative 1 percent innovation lowers the level of GNP by over 1 percent over any foreseeable horizon. Moreover, when the nonparametric procedure suggested by John Cochrane (1986) is applied to this data, the same finding is obtained. Authors using unobserved components models (Mark Watson, 1986; Peter Clark, 1987) have typically found less persistence. Even so, the long-run impact of a shock to GNP is usually estimated to be

[†] *Discussants*: James H. Stock, Harvard University; Kenneth D. West, Princeton University; Stephen P. Zeldes, University of Pennsylvania.

* Princeton University, Princeton N.J. 08544 and Harvard University, Cambridge, MA 02138, respectively. We are grateful to James Poterba for helpful comments.

[1] This dynamic pattern also revealed itself in explanations of the business cycle. Robert Barro and Mark Rush (1980), for example, estimated that unanticipated money has its greatest effect with a lag of 3 quarters, but has no effect after 10 quarters.

about 0.6. Hence, almost all recent studies have rejected the traditional view that output shocks have little or no permanent effect.

I. Two Concepts of Persistence

What is persistence? For some purposes, a shock to an economy may be considered persistent if it lasts for more than one period. Here, however, we take persistence as meaning "continuing for a long time into the future." More formally, suppose that the change in log of GNP is a stationary process with moving average representation

$$\Delta Y_t = A(L)\varepsilon_t, \tag{2}$$

where $A(L)$ is an infinite polynomial in the lag operator, and ε_t is white noise. The impact of a shock in period t on the *growth rate* in period $t+k$ is A_k. The impact of the shock on the *level* of GNP in period $t+k$ is therefore $1+A_1+\dots+A_k$. The ultimate impact of the shock on the level of GNP equals the infinite sum of these moving average coefficients, which is $A(1)$. The value of $A(1)$ is the measure of persistence we used in our previous paper. For a random walk, $A(1)$ equals one; for any series stationary around a deterministic trend, $A(1)$ equals zero.

Cochrane has recently proposed another measure of persistence. His measure can be written either as a ratio of variances or as a function of autocorrelations:

$$\frac{1}{k+1}\frac{\mathrm{Var}(Y_{t+k+1}-Y_t)}{\mathrm{Var}(Y_{t+1}-Y_t)} \equiv 1+2\sum_{j=1}^{k}(1-(j/(k+1)))\rho_j \tag{3}$$

where ρ_j is the jth autocorrelation of ΔY_t. If Y_t follows a random walk, then the variance of the $(k+1)$ lagged difference is $(k+1)$ times the variance of the once-lagged difference. Hence, for a random walk, the above ratio is one. For any stationary series, the variance of the $(k+1)$ lagged difference approaches twice the variance of the series. Hence, for any stationary series, the above ratio approaches zero for large k. Cochrane therefore proposes using the limit of the variance ratio as a measure of persistence. We call this limiting variance ratio V.

For two simple cases—a stationary process and a random walk—the two measures of persistence produce the same number. More generally, however, the two measures are not the same. Define $R^2 \equiv 1-\mathrm{Var}(\varepsilon)/\mathrm{Var}(\Delta Y)$, the fraction of the variance that is predictable from knowledge of the past history of the process. Then $A(1)$ can be expressed as

$$A(1) \equiv \sqrt{V/(1-R^2)}\,. \tag{4}$$

Equation (4) shows that the square root of Cochrane's measure of persistence is a lower bound on $A(1)$. The more highly predictable is the differenced process, the greater is the disparity between the two measures.

II. Approaches to Estimating Persistence

At least three general approaches have been proposed for estimating persistence. Here we provide a brief overview.

A. *The ARMA Approach*

In our previous paper, we modeled the change in log GNP as a stationary ARMA process. That is,

$$\phi(L)\,\Delta Y_t = \theta(L)\varepsilon_t, \tag{5}$$

where $\phi(L)$ and $\theta(L)$ are finite polynomials in the lag operator. The moving average representation $A(L)$ equals $\phi(L)^{-1}\theta(L)$, from which we computed the persistence measure $A(1)$.

If Y_t is in fact stationary around a deterministic trend, then the ARMA model in equation (5) is overdifferenced. This overdifferencing induces a unit root in the moving average component $\theta(L)$. For the ARMA approach to be able to detect stationarity, it is therefore necessary to allow for the possibility of a unit MA root by including at least one moving average parameter. Long autoregressions do not provide an adequate approximation in this case. Moreover, it is important to avoid estimation of the ARMA model via quasi-maximum likelihood techniques, since these may not provide good approximations when there is a unit MA root. In our previous paper, we estimated all models up to ARMA(3,3) using exact maximum likelihood.

We concluded that an ARMA(2,2) model well approximates the postwar, quarterly, real GNP growth. The ARMA(2,2) model and most of the other models produce similar impulse response functions, $A(L)$. After a unit shock, the maximum impact on the level of real GNP is about 1.6 after a few quarters, followed by a slight (and perhaps insignificant) decline. The long-run impact $A(1)$ was estimated to be about 1.5.

B. *The Nonparametric Approach*

One can estimate the persistence measure V very simply by replacing the population autocorrelations in equation (3) with the sample autocorrelations.[2] The estimator is

$$(6)\qquad \hat{V}^k \equiv 1 + 2\sum_{j=1}^{k}\left(1-\frac{j}{k+1}\right)\hat{\rho}_j.$$

As long as k increases with the sample size, this estimator consistently estimates V.

It is also possible to compute nonparametrically an approximate estimate of $A(1)$, called $\hat{A}^k(1)$, as

$$(7)\qquad \hat{A}^k(1) \equiv \sqrt{\hat{V}^k/\left(1-\hat{\rho}_1^2\right)}\,.$$

The estimate of $A(1)$ is computed by replacing the R^2 in equation (4) with the square of the first autocorrelation. Since ρ_1^2 is an underestimate of R^2, except for an AR(1) process, this estimate tends to understate $A(1)$.

In any given sample, it is of course necessary to choose k, the number of autocorrelations to include. Including too few autocorrelations may obscure trend reversion manifested in higher autocorrelations. Including too many autocorrelations may tend to find excessive trend reversion, since as k approaches the sample size T, the estimator approaches zero. Since the sample mean has been removed from the data, $\hat{V}^k$ is identically zero at $k=T-1$. Hence, while large k appears preferable, k must be small relative to the sample size.

TABLE 1—PERSISTENCE ESTIMATES AND THE WINDOW SIZE

Window Size (k)	Persistence Measure V		$A(1)$	
A. True Process is a Random Walk				
10	0.90	(0.29)	0.94	(0.15)
20	0.83	(0.39)	0.89	(0.21)
30	0.76	(0.46)	0.84	(0.25)
40	0.67	(0.48)	0.78	(0.27)
50	0.59	(0.49)	0.72	(0.29)
60	0.51	(0.48)	0.65	(0.30)
75	0.40	(0.43)	0.56	(0.30)
100	0.21	(0.30)	0.38	(0.28)
B. True Process is $Y_t = 1.34Y_{t-1} - 0.42Y_{t-2}$				
10	1.32	(0.38)	1.24	(0.20)
20	0.83	(0.32)	0.97	(0.20)
30	0.57	(0.27)	0.80	(0.19)
40	0.43	(0.23)	0.69	(0.19)
50	0.34	(0.21)	0.60	(0.19)
60	0.27	(0.19)	0.53	(0.19)
75	0.21	(0.17)	0.45	(0.20)
100	0.11	(0.14)	0.31	(0.21)

Note: This table presents the results of a Monte Carlo experiment. It displays the mean of the persistence estimate and, in parentheses, the standard deviation of the estimates. These results are based on a sample of 130 and 500 replications.

To examine more precisely the choice of window size (k), we have performed a small Monte Carlo experiment. The true process is, alternatively, a random walk with drift or a stationary AR(2) process around a deterministic time trend. The parameters for the AR(2) process are those estimated by Blanchard for detrended real GNP (equation (1)). There are 500 replications, and each has 130 observations, which is the number of postwar quarterly observations we use below. Table 1 reports the mean of the two persistence estimates for various k, as well as the standard deviation of the estimates.

[2] We compute the jth sample autocovariance as the sum of the $T-j$ cross products divided by $T-j$. This computation does not guarantee that the estimate of the variance ratio is positive, however. Dividing the $T-j$ cross products by T would guarantee a positive estimate. In practice, as long as k is small relative to the sample size, the difference is not important.

TABLE 2—PERSISTENCE OF POSTWAR QUARTERLY REAL GNP

Window Size (k)	Estimate of V	Estimate of $A(1)$
10	1.43 (0.48)	1.27
20	1.09 (0.50)	1.11
30	1.02 (0.57)	1.07
40	0.85 (0.55)	0.98
50	0.64 (0.46)	0.85
60	0.57 (0.45)	0.80

Note: Asymptotic standard errors are shown in parentheses. See our earlier paper (1987).

The results in Table 1 show clearly that the window size k must be small relative to the sample size. For $k=50$, the mean estimate of V for the random walk is 0.59, in contrast to a population parameter of unity. There is thus severe downward bias for large k.[3]

The results in Table 1 also show how difficult it is to distinguish between the two representations on the basis of these nonparametric persistence estimates. For $k=40$, the mean estimate of V is 0.67 for a random walk and 0.43 for the stationary AR(2). Moreover, the associated standard deviations are substantial. One cannot make inferences from these persistence estimates with great confidence.

In Table 2 we report $\hat{V}^k$ and $\hat{A}^k(1)$ for the log of quarterly real GNP from 1952:2 to 1984:3 for various values of k. The results are consistent with those obtained with the ARMA approach. A comparison of the estimates in Table 2 with the Monte Carlo results in Table 1 shows that real GNP appears more persistent than the stationary AR(2) process in equation (1). For $k=40$, the estimate of V is 0.85, which is about two standard deviations above the 0.43 mean of the AR(2) process. Similarly, the estimate of $A(1)$ of 0.98 is about one standard deviation above the 0.69 mean of the AR(2). Indeed, the estimates for real GNP slightly exceed the mean persistence estimates for a random walk.

C. *The Unobserved Components Approach*

An alternative way to model output is as the sum of two or more components. These components are not directly observed, but their relative importance, and the implications for persistence, are inferred from the time series behavior for output. Authors such as Stephen Beveridge and Charles Nelson (1981), Watson, and Clark have used this approach.

Obviously some restrictions must be imposed on the components if they are to be identifiable from a single time-series for output. Beveridge and Nelson proposed a decomposition into two components, a random walk and a stationary component, whose innovations are perfectly correlated. For any output process, these components are just identified.

The two measures of persistence can be interpreted in terms of the Beveridge-Nelson decomposition. Cochrane points out that the measure V is the variance of the change in the random walk component divided by the variance of the total change in output. It follows from equation (4) that the measure $A(1)$ is the standard deviation of the change in the random walk component divided by the standard deviation of the univariate innovation to output. It is tempting to conclude that V and $A(1)$ convey information about the relative importance of "trend" and "cycle" components in output growth. But of course, one usually thinks of trend and cycle as having a low or zero correlation, while the Beveridge-Nelson components are *perfectly* correlated.

An alternative type of unobserved components model, estimated by Watson and Clark, expresses output as the sum of two *independent* processes. The independence of the two components implies that the measure V can be written as a weighted average of the equivalent measures for the two components. Formally, if V_1 and V_2 are the persistence of components 1 and 2, and the changes in these components have variances

[3] It appears that, for the random walk, this bias can be approximately corrected by multiplying by $T/(T-k)$.

σ_1^2 and σ_2^2, then

$$V = \lambda V_1 + (1-\lambda)V_2 \tag{8}$$

where $\lambda = \sigma_1^2/(\sigma_1^2+\sigma_2^2) = \sigma_1^2/\mathrm{Var}(\Delta Y)$.

Watson and Clark assume that the first component is a random walk and the second component is a stationary AR(2), so that $V_1 = 1$ and $V_2 = 0$. Since the sum of a random walk and an AR(2) process can be represented as an ARMA(2,2) process in the growth rates, the Watson-Clark model can be viewed as imposing restrictions on this ARMA model. As in the Beveridge-Nelson decomposition, V equals the share of the random walk in output variability. Note, however, that V and $A(1)$ must now be no greater than unity to be consistent with the model. Watson and Clark thus rule out highly persistent processes a priori.

Watson and Clark estimate the parameters of the unobserved components model via maximum likelihood and then infer V and $A(1)$. Watson, for example, finds that $V = 0.36$ and $A(1) = 0.57$, suggesting that both the permanent and transitory components of GNP are important. Clark's results are comparable.

III. Using the Unemployment Rate to Measure the Business Cycle

The unobserved components models, in contrast to the unrestricted ARMA models or the nonparametric approach, suggest that there is a substantial temporary component to real GNP. The results with these models, therefore, are easier to reconcile with standard theories of the business cycle. One is tempted to interpret the permanent component as the natural rate of output, and the temporary component as the deviation of output from the natural rate. Indeed, the appeal of the restrictions imposed by Watson and Clark comes largely from the long tradition of separating issues of long-run growth from issues of short-run fluctuations.

The purpose of this section is to examine whether one can decompose output fluctuations into a transitory component associated with the business cycle and a persistent component unrelated to the business cycle. The decompositions of Watson and Clark are completely univariate. In contrast, we use the unemployment rate as a measure of the business cycle.[4]

TABLE 3—PERSISTENCE OF POSTWAR QUARTERLY UNEMPLOYMENT RATE

Window Size (k)	Estimate of V	Estimate of $A(1)$
10	1.42 (0.47)	1.53
20	0.81 (0.38)	1.15
30	0.67 (0.38)	1.05
40	0.49 (0.32)	0.90
50	0.31 (0.23)	0.72
60	0.23 (0.18)	0.62

Note: See Table 2.

We begin by examining the persistence of fluctuations in the rate of unemployment. Table 3 presents estimates of V and $A(1)$ computed as in equations (6) and (7) using quarterly data from 1952:2 to 1984:3. The results for V indicate that shocks to unemployment are no more persistent than the AR(2) process simulated in Table 1. Yet, since ρ_1 is much larger for unemployment, the estimates of $A(1)$ are larger.

We next use the unemployment rate to separate the cyclical and trend components of real GNP. As in the unobserved components models discussed above, we make the identifying assumption that these two components are uncorrelated. In contrast to these models, we do *not* assume the two components are unobserved. Instead, we assume that the cyclical component is that part of GNP correlated with unemployment at leads and lags, while the trend component is that part of GNP uncorrelated with unemployment. Hence, one might refer to our decomposition as an "observed components model."

We accomplish this decomposition by regressing the change in log real GNP on eight leads, eight lags, and the contemporaneous detrended unemployment rate. This regression yields an R^2 of about 0.6. We take the

[4] George Evans (1986) also studies persistence in a bivariate model of output and unemployment. His use of restricted and unrestricted vector autoregressions is very different from the approach taken here.

TABLE 4—PERSISTENCE OF THE CYCLICAL AND TREND COMPONENTS OF REAL GNP

Window Size (k)	Estimate of V	Estimate of $A(1)$
A. The Cyclical Component		
10	1.92 (0.64)	1.87
20	1.25 (0.58)	1.51
30	1.09 (0.61)	1.41
40	0.82 (0.53)	1.23
50	0.55 (0.40)	1.00
60	0.40 (0.32)	0.85
B. The Trend Component		
10	0.39 (0.13)	0.64
20	0.38 (0.17)	0.63
30	0.35 (0.19)	0.61
40	0.31 (0.20)	0.58
50	0.29 (0.21)	0.56
60	0.31 (0.24)	0.57

Note: See Table 2.

fitted value of this regression as a measure of the change in the cyclical component of real GNP. The residual is a measure of the change in the trend component.

Table 4 presents estimates of the persistence of the cyclical and trend components of real GNP. The persistence of the two components of GNP appears roughly equal. Indeed, to the extent that the two components differ, the cyclical component of GNP appears somewhat more persistent.

These results call into question the identifying assumptions underlying unobserved components models. These models are based on the premise that fluctuations in output associated with the business cycle are less persistent than other fluctuations in output. Yet when the unemployment rate is used as a measure of the business cycle, it is hard to find a significant difference in the persistence of these two components.

If one imposes some restrictions on the regression decomposing GNP growth, it is possible to find some evidence for the view that the cyclical component is less persistent than the trend component. In particular, including fewer leads and lags of the unemployment rate, or including the unemployment rate only in differenced form, tends to make the fitted value of the regression somewhat less persistent. In these cases, the estimated V for large k is smaller for the cyclical component, while the estimated $A(1)$ remains larger for this component. If there were reason to believe these constraints a priori, then the sort of decomposition suggested here might be interpreted as providing weak evidence for the use of unobserved components models.

IV. Conclusions

Not very long ago, it was widely believed that shocks to output almost fully dissipate in four or five years. The research summarized here has challenged that stylized fact. Fluctuations in output appear far more persistent. Indeed, it hard to reject the view that postwar real GNP is as persistent as a random walk with drift.

Of course, in any finite set of data, it is impossible to reject the view that output eventually returns to a deterministic time trend. The sort of evidence presented here can only demonstrate that such trend reversion does not occur as quickly as was once believed.

Some economists have dismissed the recent findings of persistence in real GNP because these findings do not distinguish the business cycle from other fluctuations. We have attempted to make such a distinction by using the unemployment rate as a measure of the business cycle. In contrast to what many economists have assumed, fluctuations associated with the business cycle are not obviously more trend-reverting than other fluctuations in output.

REFERENCES

Barro, Robert J. and Rush, Mark, "Unanticipated Money and Economic Activity," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Beveridge, Stephen and Nelson, Charles R., "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle,'" *Journal of Monetary Economics*, March 1981, 7, 151–74.

Blanchard, Olivier J., "What is Left of the Multiplier-Accelerator?," *American Economic Review Proceedings*, May 1981, *71*, 150–54.

Campbell, John Y. and Mankiw, N. Gregory, "Are Output Fluctuations Transitory?," *Quarterly Journal of Economics*, forthcoming 1987.

Clark, Peter K., "The Cyclical Component of U.S. Economic Activity," *Quarterly Journal of Economics*, forthcoming 1987.

Cochrane, John H., "How Big is the Random Walk in GNP?," University of Chicago, 1986.

Dickey, David A. and Fuller, Wayne A. "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica*, July 1981, *42*, 1057–72.

Evans, George W., "Output and Unemployment Dynamics in the United States: 1950–85," Stanford University, 1986.

Kydland, Finn E. and Prescott, Edward C., "A Competitive Theory of Fluctuations and the Feasibility and Desirability of Stabilization Policy," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Nelson, Charles R. and Plosser, Charles I., "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Watson, Mark W., "Univariate Detrending Methods with Stochastic Trends," *Journal of Monetary Economics*, July 1986, *18*, 1–27.

Are Cyclical Fluctuation in Productivity Due More to Supply Shocks or Demand Shocks?

By MATTHEW D. SHAPIRO*

Productivity plays a central role in the business cycle. Measured productivity varies positively with output. The procyclicality of productivity is a focus of recent debates over the sources of economic fluctuations. Real business cycle theories take shocks in productivity as a source of business cycles.[1] (See Finn Kydland and Edward Prescott, 1982; John Long and Charles Plosser, 1983; Prescott, 1986; and myself, 1986). These theories explain the joint movement of output and measured productivity virtually by definition.

Keynesian theories, on the other hand, attribute the business cycle to demand shocks. Such shocks include changes in fiscal policy, taste, velocity, and autonomous investment or animal spirits. Keynesian theories must explain the procyclical fluctuation in productivity. The sticky wage version of the Keynesian model found in the *General Theory* and more recently in models of overlapping contracts (see Stanley Fischer, 1977, for example) do not explain procyclical productivity. In these models, firms are always on their demand for labor schedules. Hence, shocks to output demand would reduce the marginal product of labor and lead to countercyclical productivity. The countercyclicality of productivity in sticky wage models is the dual of the countercyclicality of real wages. That unsatisfactory feature of the sticky wage model has been widely discussed since the Dunlop-Tarshis-Keynes debate.

The Keynesian explanation for procyclical productivity is that firms do not adjust their labor input in light of short-run fluctuation in demand because it is too costly to do so. This leads to "labor hoarding," or short-run "off the production function" behavior. Such behavior on the part of firms need not be irrational, but can be motivated by complications in the production technology (costs of adjustment, overhead labor) not captured in standard short-run production functions. Such behavior also provides part of the theoretical underpinnings of Okun's Law. (See Rudiger Dornbusch and Fischer, 1981, pp. 368–71, for a Keynesian account of procyclical productivity.)

Robert Hall, in an important series of papers (1986a, b, c), reinterprets the finding that productivity is procyclical. If a demand shock can lead to an increase in output with little increase in input, then marginal cost must be low. Competitive firms with the ability to increase output with little increase in input would cut price. Demand would increase and hence attenuate the procyclicality of measured productivity. Hence, Hall interprets the procyclicality of productivity as evidence that firms behave monopolistically and that they have consistent excess capacity. Hall's explanation is within the standard Keynesian tradition discussed in the previous paragraph, although it is important and distinct in its implications for market structure. Yet in either Hall's or the textbook Keynesian model, cyclical fluctuations in productivity arise from shocks to aggregate demand rather than shocks to true productivity.

In this paper, I attempt to test whether observed fluctuations in productivity are more from supply (real business cycles) or

*Cowles Foundation for Research in Economics, Yale University, Box 2125 Yale Station, New Haven, CT 06520 and National Bureau of Economic Research. This work was completed while I was an NBER-Olin Fellow. Financial support from National Science Foundation grant no. SES-8521791 and research assistance by Mico Loretan are also gratefully acknowledged.

[1] Given the finding that most of the variance of output changes is explained by a permanent component (see John Campbell and N. Gregory Mankiw, 1987), it is necessary to clarify what is meant by the business *cycle*. Here, a variable is said to be *cyclical* if it moves positively with innovations in aggregate output.

from demand (the Keynesian theory). To do so, I appeal to data on product and factor prices. Prices should provide an independent indication of the source of the productivity fluctuations. Implementations of real business cycle models have been criticized for neglecting their predictions about factor prices despite their strong implications for them (see Lawrence Summers, 1986). In this paper, I ask whether the observed fluctuations in factor prices are consistent with hypothesis that measured productivity shocks are true productivity shocks. Further, I ask whether departures from the predicted joint movement of measured productivity and factor prices are consistent with Keynesian alternatives.

I. Productivity and Prices

Productivity is measured here as the percent change of the residual in the value-added production function. Consider a constant returns to scale production function with Hicks-neutral technological progress. Output, Y_t, is a function of labor, N_t, and capital, K_t:

$$Y_t = f(N_t, K_t) E_t^*. \tag{1}$$

The level of the true productivity shock is denoted E_t^*. Robert Solow (1957) shows that the percent change in the shock E_t^* (denoted $\Delta\varepsilon_t$) can be measured from observed data. Taking logarithmic time derivatives of (1), setting the marginal product of labor equal to the real wage, and applying Euler's Law for linearly homogenous functions yields Solow's famous residual

$$\Delta\varepsilon_t = (\Delta y_t - \Delta k_t) - \alpha_t(\Delta n_t - \Delta k_t). \tag{2}$$

The $\alpha_t = W_t N_t / P_t Y_t$ is the share of labor income in nominal output (W_t is the wage and P_t the price level) and the time derivative of the logarithm of a variable Z_t is denoted Δz_t. In the empirical work, these are approximated as logarithmic differences. Solow's residual is a measure of the percent change in total factor productivity for any constant returns to scale technology; for it to be a valid measure, only labor need be paid its marginal product.

The aim of this paper is to evaluate whether the measured Solow residual is a true shift in the production function or whether it has a demand component as Keynesian theories suggest. In the next section, I outline how data on prices help distinguish between these competing hypotheses. Before considering the precise implications of the competing hypotheses, it is worthwhile to examine the basic comovements of prices and measured productivity.

Table 1 gives the correlation of the Solow residual with rates of change of aggregate GNP, of prices, and of wages for industries in the U.S. economy.[2] The first column gives the correlation of the Solow residual with aggregate GNP growth, the second column gives the correlation with the growth in the industry price divided by the GNP deflator, and the third column gives the correlation with growth in the real wage (compensation per man-hour divided by industry price). The first column indicates that measured productivity growth varies positively with aggregate output growth. The procyclicality is particularly strong in aggregate and in manufacturing. Construction is an interesting exception. Even though its output moves strongly with aggregate output, its productivity is acyclical.

For virtually every industry, and for the aggregate economy, the changes in measured productivity are negatively correlated with real price growth. This finding is closely related to Hendrik Houthakker's (1979) that price growth and output growth are negatively correlated in industry data. The negative correlation of price growth and the Solow residual provides some evidence that the industry level shocks to productivity are shocks

[2] The data are from the annual U.S. National Income and Product Accounts from 1950 to 1985; they are revised as of July 1986. Data on output, deflators, wages,and hours are taken from section six of the NIPA (see *Survey of Current Business*, July 1986, for example). The capital stock are from the revised industry level data (see *Survey of Current Business*, August 1986). See my 1987 paper for details about the data.

TABLE 1—CORRELATIONS WITH SOLOW RESIDUAL

	Correlation of $\Delta\varepsilon$ with Change in		
	Aggregate GNP	Real Price	Real Wage
Private Industry	0.79	−0.18	0.27
Agriculture	−0.31	−0.27	0.22
Mining	0.62	−0.51	0.48
Construction	0.06	−0.84	0.78
Manufacturing	0.77	−0.46	0.31
Durables	0.76	−0.37	0.23
Lumber	0.10	−0.53	0.53
Furniture	0.37	−0.73	0.73
Stone, Clay, Glass	0.49	0.02	−0.07
Primary Metals	0.72	−0.24	0.45
Fabricated Metals	0.57	−0.61	0.51
Machinery	0.46	−0.71	0.78
Electric Eq.	0.48	−0.44	0.37
Motor Vehicles	0.49	−0.06	0.27
Other Trans. Eq.	−0.13	−0.40	0.47
Instruments	0.38	−0.41	0.32
Misc. Mfging.	0.17	−0.67	0.56
Nondurables	0.45	−0.55	0.40
Food	0.05	−0.63	0.61
Tobacco	0.28	−0.58	0.58
Textiles	0.08	−0.68	0.76
Apparel	0.34	−0.69	0.57
Paper	0.46	−0.72	0.76
Printing	0.18	−0.58	0.64
Chemicals	0.43	−0.60	0.50
Petroleum	0.37	−0.18	0.06
Rubber	0.54	−0.37	0.38
Leather	0.30	−0.61	0.57
Transportation	0.68	−0.57	0.59
Railroads	0.71	−0.54	0.60
Local Transport	0.14	−0.78	0.76
Trucking	0.52	−0.80	0.91
Water Transport	0.39	−0.79	0.82
Air Transport	0.61	−0.45	0.55
Communications	0.21	−0.11	0.55
Electricity, Gas	0.37	−0.36	0.41
Wholesale Trade	0.36	−0.49	0.56
Retail Trade	0.47	−0.63	0.68
Finance	0.27	0.13	−0.20
Services	0.15	−0.49	0.50

to supply. Nonetheless, this evidence is only suggestive. Keynesian models with sticky prices have no restriction in general on the price-productivity correlation. Moreover, it is difficult to interpret the correlation if there are aggregate shocks to true productivity. If a productivity shock hits all industries equally, it may not change relative output prices. Hence, the price-productivity correlations, although they suggest that idiosyncratic supply shocks are important, are not helpful in studying the sources of cyclical fluctuation in industry productivity.

Fluctuations in the real (product) wage should, on the other hand, shed light on whether there are aggregate shocks to productivity. Specifically, real wages should increase if productivity increases either idiosyncratically or in aggregate. In Table 1, the correlation of real wage growth and the Solow residual is almost always positive. Hence, the data admit the possibility that the observed fluctuations in productivity are indeed from supply. In the next section, I outline a dual approach to measuring factor productivity that uses the factor price data. There is a price-based measure of productivity that should be identical to the Solow residual if the measured Solow residual is a true shock to the production function. In the subsequent section, I apply these measurements to the data to see whether movements in the two measures are similar. I also study whether their deviations are cyclical as Keynesian alternatives suggest.

II. Dual Measurement of Productivity

Under almost the same conditions that Solow uses to derive his famous residual, it is possible to derive an alternative one based on factor prices.

If a firm has a constant returns to scale technology such as that given in equation (1), it will have a cost function $C(\cdot)$ of the following form:

$$(3)\quad C(Y_t, W_t, R_t) = g(W_t, R_t)Y_t/E_t^*.$$

Here, R_t is the rental rate of capital. The function $g(\cdot)$ is, of course, linearly homogenous. Shepherd's lemma gives conditional factor demand equation. Defining marginal cost as X_t,

$$(4)\quad X_t = C_Y(Y_t, W_t, R_t) = g(W_t, R_t)/E_t^*.$$

To derive an expression for the productivity shock in terms of the dual, I follow identical steps as with the production function.

Logarithmic differentiation of (4) yields

$$(5)\quad \Delta x_t = \frac{g_W(W_t, R_t)\dfrac{dW}{dt}}{g(W_t, R_t)} + \frac{g_R(W_t, R_t)\dfrac{dR}{dt}}{g(W_t, R_t)} - \Delta\varepsilon_t^*$$

where Δx_t is the percent change in the marginal cost. Shephard's lemma implies that

$$(6)\qquad g_W(\cdot) = L_t E_t^*/Y_t;$$

$$(7)\qquad g_R(\cdot) = K_t E_t^*/Y_t.$$

Using Euler's Law, substituting into equation (5), and setting price equal to marginal cost yields the standard pricing equation. Denote the percent change in the wage as Δw_t and the percent change in the rental cost of capital as Δr_t. Under competition, the price growth (Δp_t) is given by

$$(8)\quad \Delta p_t = \alpha_t \Delta w_t + (1-\alpha_t)\Delta r_t - \Delta\varepsilon_t^*,$$

where $\Delta\varepsilon_t^*$ is total factor productivity growth. Rearranging yields

$$(9)\quad \Delta\varepsilon_t^p = \alpha_t(\Delta w_t - \Delta p_t) + (1-\alpha_t)(\Delta r_t - \Delta p_t),$$

where the price-based measurement of the productivity shock is labeled $\Delta\varepsilon_t^p$ to distinguish it from the Solow residual $\Delta\varepsilon_t$ defined in (2).

This productivity residual is derived under almost as general assumptions as is Solow's. The only extra assumption made is that capital is paid its marginal product within the period. This assumption is clearly unrealistic because of costs of adjustment and time to build. It can be relaxed at some cost in generality. Suppose that changes in capital must be decided at least one period in advance. Then marginal cost becomes simply

$$(10)\qquad X_t = w_t(\partial N_t/\partial Y_t).$$

The derivative of labor with respect to output will be a function of the capital stock, the level of output, and the productivity shock. Equation (10) can be logarithmically differentiated and an expression similar to (9) can be derived. Unfortunately, to obtain an analytic expression, it is necessary to parameterize the production function. The ensuing result is not as general as (9). Suppose that the production function is constant elasticity of substitution (*CES*) so

$$(11)\quad Y_t = [(1-\alpha)K_t^\rho + \alpha N_t^\rho]^{1/\rho} E_t^*,$$

where α is a distribution parameter and ρ is a parameter such that $1/(1-\rho)$ is the elasticity of substitution between capital and labor. For the Cobb-Douglas case $(\rho = 0)$, the dual measure of productivity growth becomes

$$(12)\quad \Delta\varepsilon_t^p = \alpha(\Delta w_t - \Delta p_t) + (1-\alpha)(\Delta y_t - \Delta k_t).$$

In the general *CES* case, it is

$$(13)\ \Delta\varepsilon_t^p = [\alpha(\Delta w_t - \Delta p_t) + (1-\alpha)(1-\rho) \times (\Delta y_t - \Delta k_t)]/(1-\rho(1-\alpha)).$$

These expressions are similar to (9). They are still importantly dependent on the real wage. Instead of the term in the cost of capital, they have terms that reflects the marginal product of capital in terms of quantities. These capture the short-run increasing marginal cost due to the fixity of capital. Kydland and Prescott assume that the change in the capital stock is predetermined and that the production function is Cobb-Douglas, so it seems appropriate to consider that case here.[3]

II. Primal versus Dual Productivity: Empirical Findings

There are two ways to measure productivity change: the standard, output-based mea-

[3] They also consider inventories, which are ignored here.

sure, and the dual, price-based measure. Under the null hypothesis that the measured changes in productivity are true changes in productivity, these measures should be identical except for measurement errors or specification errors from incorrect parameterization of the production function. In this section, I compare these two measures. Moreover, I study whether the two measures depart from each other in a way that would be predicted by a Keynesian alternative. Under such an alternative, the Solow residual moves independently with aggregate demand. Cyclical fluctuations in quantity-based productivity occur because firms hoard labor or are off their production functions. Under the Keynesian alternative, the cyclical fluctuations in the quantity-based measure of productivity have nothing to do with the true productivity of the factors of production, so factor prices should not move in response to these cyclical fluctuations. Hence, the deviation of the quantity-based and price-based measures should be cyclical under the Keynesian alternative.

Consider first a regression of the Solow residual $\Delta\varepsilon_t$ on the dual residual $\Delta\varepsilon_t^p$ and a constant. Under the null hypothesis that the two productivity measurements are equal, the slope coefficient and the R^2 should both equal one. For aggregate manufacturing, using the unrealistic specification that capital is flexible (equation (9)), the estimates are as follows:[4]

$$(14)\quad \Delta\varepsilon_t = \underset{(0.95)}{0.03} + \underset{(0.35)}{0.79}\Delta\varepsilon_t^p,$$

$$SEE = 3.01,\quad D\text{-}W = 1.95,\quad R^2 = 0.13.$$

Even in this specification, the hypothesis that the slope coefficient is one cannot be rejected although the fraction of variance explained is very low. Ignoring the short-run fixity of capital understates the variability of marginal cost. In the Cobb-Douglas specification (equation (12)), the results are as follows:[5]

$$(15)\quad \Delta\varepsilon_t = -\underset{(0.34)}{0.31} + \underset{(0.11)}{1.10}\Delta\varepsilon_t^p,$$

$$SEE = 1.58,\quad D\text{-}W = 2.03,\quad R^2 = 0.76.$$

The slope coefficient is close to one and fairly tightly estimated. Moreover, variation in the factor prices explains about three-fourths the variance in the quantity-based measure. Finally, in a *CES* specification with ρ constrained to equal -1 (elasticity of substitution equal to 0.5), the estimates are

$$(16)\quad \Delta\varepsilon_t = \underset{(0.22)}{0.54} + \underset{(0.07)}{0.88}\Delta\varepsilon_t^p,$$

$$SEE = 1.23,\quad D\text{-}W = 1.98,\quad R^2 = 0.86.$$

Again the slope coefficient is precisely estimated to be close to one and the R^2 is very high. Therefore, in the specifications where the short-run fixity of capital is taken into account, the two measures of productivity appear to be very similar.

Now consider the prediction of a Keynesian alternative where movements in measured Solow residuals are accounted for by movements in demand. This alternative can be tested directly by including a measure of demand, say the growth rate of aggregate GNP. We know from Table 1 that a regression of GNP growth alone on manufacturing productivity explains about half of the variance of measured productivity. Because the Cobb-Douglas and the case with lower elasticity of substitution yield similar results, only the results for Cobb-Douglas are presented. They are as follows:

$$(17)\quad \Delta\varepsilon_t = \underset{(0.42)}{-0.63} + \underset{(0.18)}{0.92}\Delta\varepsilon_t^p + \underset{(0.17)}{0.21}\Delta GNP_t,$$

$$SEE = 1.57,\quad D\text{-}W = 1.86,\quad R^2 = 0.77.$$

[4] The rental price of capital is calculated as $R_t = (\delta + v_t)q_t(1 - z_t\tau_t - \omega_t)/(1 - \tau_t)$, where δ is the average depreciation rate (0.125 for manufacturing), v_t is the required rate of return (measured by the dividend-price ratio on the Standard and Poor's Composite), z_t in the present discounted value of depreciation allowances, ω_t is the investment tax credit rate, and τ_t is the corporate profits tax rate. The variables z_t and ω_t are computed by DRI. Note that if this measure of profits were found in the national accounts, which of course it is not, equation (14) would be tautologous. As Hall (1986c) stresses, tests of the model that factors earn their marginal products depend critically on the measurement of profits.

[5] In the Cobb-Douglas estimates of $\Delta\varepsilon_t^p$, α is estimated as its average value.

TABLE 2—REGRESSION OF THE SOLOW RESIDUAL: COBB-DOUGLAS CASE

	Slope Coefficient in Regression Equation (15)	Fraction of Variance in $\Delta\varepsilon$ (equation (17)) Explained by	
		$\Delta\varepsilon^p$	ΔGNP
Private Industry	0.98	0.83	0.00
Agriculture	1.02	0.93	0.01
Mining	0.69[a]	0.74	0.02
Construction	1.01	0.90	0.00
Manufacturing	1.10	0.76	0.01
Durables	1.16	0.71	0.03
Lumber	0.84	0.71	0.06
Furniture	0.85	0.75	0.02
Stone, Clay, Glass	1.33	0.64	0.00
Primary Metals	1.34[a]	0.84	0.01
Fabricated Metals	0.78[a]	0.64	0.06
Machinery	1.07	0.89	0.00
Electric Eq.	1.10	0.57	0.01
Motor Vehicles	1.33[a]	0.84	0.01
Other Trans. Eq.	1.05	0.35	0.04
Instruments	0.97	0.51	0.03
Misc. Mfging.	1.13	0.66	0.00
Nondurables	0.97	0.72	0.01
Food	0.70[a]	0.65	0.02
Tobacco	0.98	0.98	0.00
Textiles	0.99	0.78	0.01
Apparel	0.63[a]	0.45	0.00
Paper	1.11	0.87	0.01
Printing	1.03	0.71	0.01
Chemicals	1.23[a]	0.84	0.01
Petroleum	0.64[a]	0.41	0.07
Rubber	0.97	0.57	0.06
Leather	0.74	0.48	0.01
Transportation	1.05	0.82	0.02
Railroads	1.11	0.74	0.04
Local Transport	0.72[a]	0.79	0.01
Trucking	0.94	0.94	0.00
Water Transport	0.98	0.83	0.00
Air Transport	0.99	0.72	0.11
Communications	0.79	0.58	0.04
Electricity, Gas	1.05	0.82	0.00
Wholesale Trade	1.08	0.76	0.03
Retail Trade	0.86	0.82	0.00
Finance	1.09[a]	0.95	0.00
Services	0.84	0.76	0.03

[a] Denotes significantly different from one at the 5 percent level. See text for definitions of equations (15) and (17).

The addition of GNP growth adds almost nothing to the explanatory power of the equation. Moreover, the null hypothesis that the coefficient on $\Delta\varepsilon_t^p$ is one and the coefficient on ΔGNP_t is zero cannot be rejected (the $F(2,33)=1.25$ statistic has marginal significance 0.30).

In the Cobb-Douglas case, the difference of $\Delta\varepsilon_t$ and $\Delta\varepsilon_t^p$ is simply $\alpha[(\Delta y_t - \Delta n_t) - (\Delta w_t - \Delta p_t)]$. Hence, imposing the restriction that the slope coefficient in equation (17) is one yields a regression of the difference of labor productivity and real wage growth on aggregate output growth. This difference, as equation (17) indicates, is not cyclical. When the restriction is imposed on equation (17), the coefficient of aggregate GNP growth remains insignificant and the equation has a R^2 of only 0.07.

Table 2 gives similar results for all the industries studied. The first column reports the estimated slope coefficient for equation (15), that is a regression of the Solow residual on the Cobb-Douglas dual residual. Many of the point estimates are close to one and precisely estimated. In about a quarter of the estimates, it is possible to reject the null hypothesis. The final two columns give the fraction of variance in $\Delta\varepsilon_t$ explained by either $\Delta\varepsilon_t^p$ or by ΔGNP_t.[6] In general, a much higher fraction is explained by the prices than by aggregate output. The result reported in equation (17) of the text holds for many industries: the deviation of the two productivity measures is acyclical.

III. Discussion

Productivity can be measured by either prices or quantities. If measured productivity is equal to true productivity, these two measures should be identical. Indeed, the two measures are very closely related: for the aggregate and for most industries, the coefficient of the dual measure in a regression of the primal is precisely estimated to be one; the fraction of variance explained by dual measure is high. Of course, the estimated R^2 is not exactly unity as the theory predicts, but the deviations from that value can easily be attributed to specification and measurement errors.

More importantly, the deviation of the two measures is not cyclical. Under the hypothesis that the coefficient of $\Delta\varepsilon_t^p$ is one in equation (17), which cannot be rejected for two-thirds of the industries, equation (17)

[6] The decomposition attributes the covariance of $\Delta\varepsilon_t^p$ and ΔGNP_t to $\Delta\varepsilon_t^p$, which is appropriate given the difficulty rejecting the hypothesis that growth in aggregate output does not enter (17).

can be interpreted as a regression of the difference of labor productivity and real wage growth rates on aggregate output growth. Keynesian theories of labor hoarding or of monopolistic excess capacity predict that measured labor productivity is procyclical. Output can increase autonomously from an increase in inputs when demand increases. Because true productivity is not increasing, there is no reason to expect an equal increase in the product wage. In the absence of a Keynesian theory that has the real wage as a better proxy for demand shocks than is aggregate GNP growth, it seems very hard to reconcile the findings in this paper with theories that make changes in conventionally measured productivity a consequence of fluctuation in aggregate demand. A possible Keynesian explanation of procyclical productivity is monopolistic theories with rent-sharing arrangements. Such theories would require marginal cost to rise precisely with the rise in observed labor productivity.

The results need to be somewhat qualified. The two measures of productivity are not exactly the same. In a few industries, the supply shock story fails importantly. More importantly, the data used in this paper are annual. Perhaps rigidities such as sticky prices and labor hoarding are confined to operate within the year. Further tests are needed on data of higher frequency. Nonetheless, if these Keynesian phenomena are indeed confined to operate over a horizon of a year, the supply shock model has explained much of the conventionally defined business cycle.

REFERENCES

Campbell, John Y. and Mankiw, N. Gregory, "Permanent and Transitory Components in Macroeconomic Fluctuations," *American Economic Review Proceedings*, May 1987, *77*, 111–17.

Dornbusch, Rudiger, and Fischer, Stanley, *Macroeconomics*, 2nd ed. New York: McGraw-Hill, 1981.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *88*, 191–205.

Hall, Robert E., (1986a) "The Relation Between Price and Marginal Cost in U.S. Industry," unpublished paper, Hoover Institution, 1986.

______, (1986b) "Market Structure and Macro Fluctuation," *Brookings Papers on Economic Activity*, 2:1986, 285–322.

______, (1986c) "Chronic Excess Capacity in U.S. Industry," NBER Working Paper No. 1973, 1986.

Houthakker, Hendrik S., "Growth and Inflation: Analysis by Industry," *Brookings Paper on Economic Activity*, 1:1979, 241–56.

Kydland, Finn E. and Prescott, Edward C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Long, John B., Jr. and Plosser, Charles I., "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

Prescott, Edward C., "Theory Ahead of Business Cycle Measurement," *Carnegie-Rochester Conference Series on Public Policy*, Autumn 1986, *25*, 11–44.

Shapiro, Matthew D., "Investment, Output, and the Cost of Capital," *Brookings Papers on Economics Activity*, 1:1986, 111–52.

______, "Measuring Market Power in U.S. Industry?," unpublished, National Bureau of Economic Research, 1987.

Solow, Robert M., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, *39*, 312–20.

Summers, Lawrence H., "Some Skeptical Observations on Real Business Cycle Theory," *Federal Reserve Bank of Minneapolis Quarterly Review*, Fall 1986, *10*, 23–26.

The Development of Keynesian Macroeconomics

By BENNETT T. MCCALLUM*

In this paper I will address the topic of this session by outlining the historical development of Keynesian macroeconomics and add a few remarks on the issues of today. My version of the story will agree with textbook accounts in some ways, and differ fairly sharply in others. Very few references will be provided in support of assertions both because space is limited and because my version is something of a "stylized history of thought." The hope is that it will be, like carefully selected "stylized facts," analytically illuminating although lacking in detail.

I. John Maynard Keynes

During the past twenty years, there has grown up a body of literature that promotes the notion that Keynes's own theorizing was vastly superior to that of the "Keynesian" variety that typified mainstream macroeconomic analysis in the 1950's and 1960's. In my opinion, the ranking implied by this literature is precisely the opposite of that which is warranted.

The foregoing contention is based on an evaluation of the contribution to business-cycle theory provided by Keynes' *General Theory* (*GT*). Any such evaluation must, it seems clear, be made in light of pre-existing theory. My own nonextensive reading of pre-*GT* writings has led me to the view that the main analytical ingredients of the *GT* were distinctly present in the writings of Alfred Marshall and his other students. In particular, Marshall (1887) and Frederick Lavington (1922) described the mechanism of cyclical fluctuations in a manner that (*i*) emphasized the sluggishness of nominal wage adjustments and (*ii*) utilized multiplier effects to explain the magnitude of departures from normality.[1] Furthermore, the idea that these fluctuations were viewed as unimportant by the pre-*GT* Cambridge writers is soundly refuted by the introductory chapter of Lavington's little book (pp. 9–12).

Of course the *GT* had an enormous influence in terms of introducing new concepts and terminology, posing new issues and puzzles, and generally redirecting economists' attention. In addition, the *GT* represented an ambitious attempt to bring the Marshallian building blocks together in the form of a detailed, rigorous, and comprehensive model that would be useful for aggregative analysis. But in this admirable attempt at formal theory, Keynes failed. His top-priority goal of articulating a model with an unemployment equilibrium—in the sense of a situation from which there is no tendency to depart—foundered on the Pigou-Patinkin real balance effect. And as a comprehensive analytical structure, the *GT* was plagued by various logical inconsistencies,[2] which were straightened out only in the more careful works of John Hicks (1937), Franco Modigliani (1944), and Don Patinkin (1956). These clarifications left the profession with the analytical structure that Keynes had evidently been seeking. But this structure owed its fundamental ideas to Marshall and other earlier

[†] *Discussants*: Robert J. Barro, University of Rochester, Robert E. Lucas, Jr., University of Chicago.

*Professor of Economics, Carnegie-Mellon University, Pittsburgh, PA 15213.

[1] See Marshall (p. 358) and Lavington (pp. 48–51 and 81–86).

[2] Numerous examples are detailed by Don Patinkin (1956; 1982).

writers, and its analytical precision to Hicks, Modigliani, and Patinkin.[3]

II. Keynesian Macroeconomics

The key characteristic of Keynesian macroeconomics that distinguishes it from Classical theory is a postulated stickiness in some nominal price that enables its value to differ, for significant spans of time, from the level that would otherwise (i.e., in the absence of this friction) be market clearing. Demand and supply quantities (defined in the absence of the friction) can differ, therefore, so fluctuations in nominal aggregate demand can be much more important for aggregate output and employment determination than under flexible-price Classical conditions.

Sluggishness of price adjustments is inherently a dynamic concept, but the refined Hicks-Modigliani-Patinkin version was, like the *GT* itself, expressed in the form of a static model. Consequently, the way in which price sluggishness had to be reflected was in the model's concept of a "short-run" equilibrium. Formally, what this amounted to was a mode of analysis centering on equilibria of a *conditional* variety: the refined *GT* model was designed to determine values of endogenous variables conditional upon "given" values of specified prices (most often, nominal wages). This is, to reiterate, the way in which the hypothesis of slow price adjustments was expressed in a framework that was formally static.

Now the object of constructing the model was to provide analytical guidance for the design of macroeconomic policy. But actual economies are dynamic, not static, so some way had to be found to relate the model to reality. One possible way of proceeding would be to choose policy actions at any point in time (say, t) by treating the current value of the sticky price (say, W_t) as historically *given* and essentially ignoring the future—since it can be attended to when it arrives. Then in period $t+1$, the wage W_{t+1} would be treated as historically given, and a new policy action chosen conditional upon its value. In this way it would be possible to use the model without ever developing any explanation for the economy's W_t values.

Of course, it is apparent that this way of proceeding would be highly undesirable. For even if W_t were actually given in t as a residue of the past, the particular value prevailing would certainly have been influenced by economic conditions of the past. The temporarily fixed price in the Keynesian model is properly viewed as a predetermined variable, not an exogenous variable. So policy actions taken in t will have effects on future prices—on W_{t+1}, W_{t+2}, etc.—and those effects are ignored in the procedure under discussion. This point is worth mentioning in our history because the procedure is a stylized version of a common method of policy analysis as actually conducted in the 1950's and 1960's. Furthermore, the efforts of many distinguished theorists were devoted to the refinement of conditional equilibrium models as recently as the late 1970's.[4]

III. Phillips Curves

Many Keynesian analysts recognized the undesirability of conditional equilibrium analysis, of course, and adopted a different approach. Rather than treating W_t as coming out of the blue, this second approach was to add to a static Keynesian model another equation or sector designed to explain movements over time[5] in the slowly adjusting price W_t. Then the model would be dynamic,

[3] Patinkin (1982) has emphasized the originality of Keynes' insight that, with fixed prices, output adjustments can provide an equilibrating mechanism. This argument holds in a clean form, however, only in a model in which interest rate adjustments are suppressed by the unsatisfactory device of treating investment as exogenous.

[4] Here reference is to the surge of interest in so-called "disequilibrium" or "fixed-price" analysis. My claim is not that distinguished theorists actually embraced the policy procedure described, but that their writings could have easily been interpreted by policymakers as providing support for such a procedure.

[5] Over actual time, not the meta-time of stability analysis such as Patinkin's (1956, pp. 152–58 and 342–51).

even if incompletely based on dynamic optimization analysis, and could be used for policy experiments that would avoid the particular difficulty described above.

Equations or sectors of this type are versions of the famous Phillips curve. As all readers know, most early formulations were severely flawed in the sense of positing adjustment procedures that involve dynamic money illusion. As Milton Friedman (1968) effectively noted, these versions carried the highly implausible implication that a society could permanently keep its real rate of output high (i.e., enrich itself in real terms) by continually printing paper money at a rapid pace. A more proper specification of the Phillips Curve, according to Friedman, would relate changes in expected *real* wages to prevailing levels of output relative to normal.

IV. Rational Expectations

Friedman's contribution improved matters considerably but not, according to Robert Lucas (1981, pp. 90–95), enough. Suppose output relative to normal is systematically related to the unexpected rate of change of some nominal variable, as Friedman's reformulation would imply. Then output could still be kept high (relative to normal) permanently if actual inflation could be kept permanently above the rate expected. Such a possibility was, moreover, permitted by Friedman's model of expectational behavior, adaptive expectations. To rule out the implausible possibility of real enrichment by monetary means, Lucas suggested adoption of the hypothesis of rational expectations—that is, the absence of any systematic relation between expectational errors and information available to agents at the time of expectation formation. This hypothesis was also necessary, Lucas indicated (p. 285), to avoid the implication of suboptimal behavior by individuals.

In addition, and as importantly, Lucas (pp. 66–89) proposed a new theory of price stickiness. Instead of some algebraic representation of price adjustments (executed by some unspecified agent) in response to excess demand, Lucas suggested a model based on information limitations faced by individual sellers. The crucially desirable feature of this new approach was its strategy of explaining incomplete price responses to monetary shocks—and thus nonzero quantity responses—in terms of choices made by optimizing agents in light of their own objectives and constraints. This strategy was adopted not for aesthetic reasons, but in order to produce a model that would be well-designed for the Keynesian objective of guiding macroeconomic policy. Such would not be possible with an algebraic price adjustment equation, for the latter would give the analyst no basis for knowing whether the relation would itself shift if policy were substantially altered—which is crucial because such a shift would invalidate his predictions about the effects of the policy change. It is necessary, according to this view, to understand the *nature* of price sluggishness to know if its quantitative manifestation will remain intact in the face of altered conditions.

V. Recent Developments

Lucas's approach gained much support during the late 1970's but today (i.e., December 29, 1986) matters are rather unsettled. A major reason for this condition is that the specific informational specification proposed in Lucas's model—which requires agents to be devoid of knowledge concerning current monetary conditions—has come to be viewed as inapplicable to today's developed economies.[6] And no other model has been devised that combines empirical accuracy with a price-adjustment sector that is derived from individuals' objectives and constraints.

Consequently, there has been a splintering of opinion, with prominent researchers promoting widely divergent strategies. One small but significant group has embraced an ultraclassical "real business cycle" position, according to which aggregative output fluctuations are induced almost entirely by technology shocks, with money-output corre-

[6] The specification is much more applicable to the economies that Keynes was concerned with in the 1920's and 1930's.

lations occurring only because the monetary system responds to these fluctuations. Most macroeconomists are highly skeptical of this position; some of my own reasons are outlined in my 1986 paper.

A more sizeable group has reacted against the postulate that sluggish price adjustments need to be explained in terms of individuals' objectives and constraints. It is better, according to this view, to use a poorly understood but empirically substantiated price-adjustment relation than to pretend—counterfactually—that all nominal adjustments take place promptly. One's econometric model will then track the data better and the adjustment relation will be unlikely to shift much or rapidly when policy changes are undertaken.

It is hard to keep from having considerable sympathy with this view. But the logic of the "Lucas critique" objection is inescapable. One possible way out of the dilemma, perhaps, is to proceed with models incorporating price adjustment equations that can be rationalized by subsidiary arguments, even though these arguments cannot clearly find expression in terms of the model's explicit taste and technology representation. This is an interpretation that can perhaps be given to John Taylor's (1979) well-known formulation, though I believe some modifications would be appropriate.

A pervasive problem in devising well-rationalized models of price stickiness and monetary effects on real variables is that taste and technology analysis (even when augmented by monopoly, asymmetric information, and insurance considerations) typically proceeds entirely in real terms. Accordingly, any such models that rationalize the predetermination of prices do so, appearances notwithstanding, in terms of real (i.e., indexed) prices and therefore fail to explain the crucial phenomena. In an attempt to remedy this weakness, I have constructed an argument that justifies the possible reinterpretation of some models of this type in terms of nominal prices (see my earlier paper, 1986). The basic idea is unimpressively simple: the benefits to an individual obtained from indexation are exceedingly small. Therefore, for small and nonongoing transactions, the tiny computational costs of expressing prices in indexed form will outweigh the benefits. For such transactions, stickiness will then pertain to nominal prices.

This last argument is not entirely immune to the Lucas critique: in a regime with more rapid inflation, the benefits from indexation may be greater. The implied model incorporates, in other words, a "rule of thumb" that would tend to be revised if placed under severe strain. But the argument does not abandon rationality as an essential ingredient. In this respect it differs from some more extreme suggestions that have recently been put forth by other writers in response to the dilemma noted above. I will conclude by briefly considering two of these other positions.

One, expressed most prominently by George Akerlof and Janet Yellen (1985), suggests that certain small departures from rational behavior on the part of individual agents will have very small effects on these individuals' utility levels. Yet if many individuals are engaged in these small departures, the aggregative consequences can be quite large. In my opinion, this argument is sensible, but in one respect misstated. The point is that if the model used in the implicit definition of "rational behavior" neglects some small computational or adjustment costs, then the agents' choices hypothesized by Akerlof and Yellen may in fact be *entirely* rational. Under this interpretation, the argument becomes rather similar to the one given two paragraphs above.

The second example is the proposed abandonment of rational expectations. Here I would emphasize that to concentrate on the question "Are expectations rational?" is to miss the true issue. Of course, there are empirical departures from the hypothesized orthogonality conditions, but can the same departures plausibly be relied upon to hold in the future? The answer is no. A better way to proceed would be to suggest that recognition of adjustment and computational costs might lead to weaker formal representations of expectational rationality —for example, that expectational errors have unconditional (but not conditional) means of zero. But such an approach would (again) even if incompletely based on dynamic optimization analysis, and could be used for policy experiments that would avoid the particular difficulty described above.

Equations or sectors of this type are versions of the famous Phillips curve. As all readers know, most early formulations were severely flawed in the sense of positing adjustment procedures that involve dynamic money illusion. As Milton Friedman (1968) effectively noted, these versions carried the highly implausible implication that a society could permanently keep its real rate of output high (i.e., enrich itself in real terms) by continually printing paper money at a rapid pace. A more proper specification of the Phillips Curve, according to Friedman, would relate changes in expected *real* wages to prevailing levels of output relative to normal.

IV. Rational Expectations

Friedman's contribution improved matters considerably but not, according to Robert Lucas (1981, pp. 90–95), enough. Suppose output relative to normal is systematically related to the unexpected rate of change of some nominal variable, as Friedman's reformulation would imply. Then output could still be kept high (relative to normal) permanently if actual inflation could be kept permanently above the rate expected. Such a possibility was, moreover, permitted by Friedman's model of expectational behavior, adaptive expectations. To rule out the implausible possibility of real enrichment by monetary means, Lucas suggested adoption of the hypothesis of rational expectations—that is, the absence of any systematic relation between expectational errors and information available to agents at the time of expectation formation. This hypothesis was also necessary, Lucas indicated (p. 285), to avoid the implication of suboptimal behavior by individuals.

In addition, and as importantly, Lucas (pp. 66–89) proposed a new theory of price stickiness. Instead of some algebraic representation of price adjustments (executed by some unspecified agent) in response to excess demand, Lucas suggested a model based on information limitations faced by individual sellers. The crucially desirable feature of this new approach was its strategy of explaining incomplete price responses to monetary shocks—and thus nonzero quantity responses—in terms of choices made by optimizing agents in light of their own objectives and constraints. This strategy was adopted not for aesthetic reasons, but in order to produce a model that would be well-designed for the Keynesian objective of guiding macroeconomic policy. Such would not be possible with an algebraic price adjustment equation, for the latter would give the analyst no basis for knowing whether the relation would itself shift if policy were substantially altered—which is crucial because such a shift would invalidate his predictions about the effects of the policy change. It is necessary, according to this view, to understand the *nature* of price sluggishness to know if its quantitative manifestation will remain intact in the face of altered conditions.

V. Recent Developments

Lucas's approach gained much support during the late 1970's but today (i.e., December 29, 1986) matters are rather unsettled. A major reason for this condition is that the specific informational specification proposed in Lucas's model—which requires agents to be devoid of knowledge concerning current monetary conditions—has come to be viewed as inapplicable to today's developed economies.[6] And no other model has been devised that combines empirical accuracy with a price-adjustment sector that is derived from individuals' objectives and constraints.

Consequently, there has been a splintering of opinion, with prominent researchers promoting widely divergent strategies. One small but significant group has embraced an ultra-classical "real business cycle" position, according to which aggregative output fluctuations are induced almost entirely by technology shocks, with money-output corre-

[6] The specification is much more applicable to the economies that Keynes was concerned with in the 1920's and 1930's.

lations occurring only because the monetary system responds to these fluctuations. Most macroeconomists are highly skeptical of this position; some of my own reasons are outlined in my 1986 paper.

A more sizeable group has reacted against the postulate that sluggish price adjustments need to be explained in terms of individuals' objectives and constraints. It is better, according to this view, to use a poorly understood but empirically substantiated price-adjustment relation than to pretend—counterfactually—that all nominal adjustments take place promptly. One's econometric model will then track the data better and the adjustment relation will be unlikely to shift much or rapidly when policy changes are undertaken.

It is hard to keep from having considerable sympathy with this view. But the logic of the "Lucas critique" objection is inescapable. One possible way out of the dilemma, perhaps, is to proceed with models incorporating price adjustment equations that can be rationalized by subsidiary arguments, even though these arguments cannot clearly find expression in terms of the model's explicit taste and technology representation. This is an interpretation that can perhaps be given to John Taylor's (1979) well-known formulation, though I believe some modifications would be appropriate.

A pervasive problem in devising well-rationalized models of price stickiness and monetary effects on real variables is that taste and technology analysis (even when augmented by monopoly, asymmetric information, and insurance considerations) typically proceeds entirely in real terms. Accordingly, any such models that rationalize the predetermination of prices do so, appearances notwithstanding, in terms of real (i.e., indexed) prices and therefore fail to explain the crucial phenomena. In an attempt to remedy this weakness, I have constructed an argument that justifies the possible reinterpretation of some models of this type in terms of nominal prices (see my earlier paper, 1986). The basic idea is unimpressively simple: the benefits to an individual obtained from indexation are exceedingly small. Therefore, for small and nonongoing transactions, the tiny computational costs of expressing prices in indexed form will outweigh the benefits. For such transactions, stickiness will then pertain to nominal prices.

This last argument is not entirely immune to the Lucas critique: in a regime with more rapid inflation, the benefits from indexation may be greater. The implied model incorporates, in other words, a "rule of thumb" that would tend to be revised if placed under severe strain. But the argument does not abandon rationality as an essential ingredient. In this respect it differs from some more extreme suggestions that have recently been put forth by other writers in response to the dilemma noted above. I will conclude by briefly considering two of these other positions.

One, expressed most prominently by George Akerlof and Janet Yellen (1985), suggests that certain small departures from rational behavior on the part of individual agents will have very small effects on these individuals' utility levels. Yet if many individuals are engaged in these small departures, the aggregative consequences can be quite large. In my opinion, this argument is sensible, but in one respect misstated. The point is that if the model used in the implicit definition of "rational behavior" neglects some small computational or adjustment costs, then the agents' choices hypothesized by Akerlof and Yellen may in fact be *entirely* rational. Under this interpretation, the argument becomes rather similar to the one given two paragraphs above.

The second example is the proposed abandonment of rational expectations. Here I would emphasize that to concentrate on the question "Are expectations rational?" is to miss the true issue. Of course, there are empirical departures from the hypothesized orthogonality conditions, but can the same departures plausibly be relied upon to hold in the future? The answer is no. A better way to proceed would be to suggest that recognition of adjustment and computational costs might lead to weaker formal representations of expectational rationality —for example, that expectational errors have unconditional (but not conditional) means of zero. But such an approach would (again) not actually represent the abandonment of rationality. A true abandonment would, in my opinion, constitute suicide for the profession.

It is necessary to stop at this point. Some readers may feel that the foregoing is not actually a history of Keynesian macroeconomics, and I would have to agree that it neglects many interesting and significant matters. But I would strenuously argue that it outlines the main developments concerning the single most important aspect of Keynesian economics—or, perhaps, macroeconomics more generally.

REFERENCES

Akerlof, George and Yellen, Janet, "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibrium?" *American Economic Review*, September 1985, *75*, 708–20.

Friedman, Milton, "The Role of Monetary Policy," *American Economic Review*, March 1968, *58*, 1–17.

Hicks, John R., "Mr. Keynes and the 'Classics': A Suggested Interpretation," *Econometrica*, April 1937, *5*, 147–59.

Lavington, Frederick, *The Trade Cycle*, London: P. S. King & Staples, 1922.

Lucas, Robert E., Jr., *Studies in Business-Cycle Theory*, Cambridge: MIT Press, 1981.

McCallum, Bennett T., "On 'Real' and 'Sticky Price' Theories of The Business Cycle," *Journal of Money, Credit, and Banking*, November 1986, *18*, 397–414.

Marshall, Alfred, "Remedies for Fluctuations of General Prices," *Contemporary Review*, March 1987, *51*, 355–75.

Modigliani, Franco, "Liquidity Preference and the Theory of Interest and Money," *Econometrica*, January 1944, *12*, 45–88.

Patinkin, Don, *Money, Interest, and Prices*, New York: Harper & Row, 1956.

______, *Anticipations of the General Theory? And Other Essays on Keynes*, Chicago: University of Chicago Press, 1982.

Taylor, John B., "Staggered Wage Setting in a Macro Model," *American Economic Review Proceedings*, May 1979, *69*, 108–13.

Keynes, Lucas, and Scientific Progress

By ALAN S. BLINDER*

In one of those marvelous coincidences of intellectual history, Robert Lucas was born the year after the publication of Keynes' *General Theory*. For the first thirty-five years of their mutual lives, the two apparently coexisted in harmony. But their relationship has been tumultuous ever since. Lucas has frequently criticized Keynesian economics as poor science; and it is precisely in that spirit that I want to address the debate today.

We all know the old joke about the professor who uses the same exam questions year after year, but changes the answers. That joke encapsulates all too well what has happened to macroeconomics these last fifteen years and seems to reflect poorly on economics as a science. Or does it? On second thought, the best answers to scientific questions *do* change as new observations are made, as new experiments are run, and as better theories are developed. The issue is whether the answers to important questions in macroeconomics have changed for good scientific reasons or for other reasons.

The joke provides the framework for my talk. I will pose eight exam questions; and for each one I will summarize the answers given by Keynes, by Lucas and his followers, and by modern Keynesians. I pick only questions that are answered differently by Keynesians and Lucasians and that are central to contemporary macroeconomic debates. The focus is on whether the Keynesian or new classical answers have greater claim to being "scientific." Each student must answer every question.

I. Are Expectations Rational?

Keynes, though no stranger to probability theory, was nonetheless unequivocal in his denial that expectations are what we now call rational:

> ...a large proportion of our positive activities depend on spontaneous optimism rather than on a mathematical expectation.... Only a little more than an expedition to the South Pole, is it based on an exact calculation of benefits to come. Thus if animal spirits are dimmed and the spontaneous optimism falters, leaving us to depend on nothing but a mathematical expectation, enterprise will fade and die....
> [1936, pp. 161–62]

That attitude left a big loose end in *The General Theory*. Business investment is supposedly driven by "the state of long-term expectations," but expectations are not pinned down by the theory, leaving substantial room for gyrations in macroeconomic activity driven by autonomous changes in animal spirits. That hardly constitutes a tight scientific theory; but Keynes was probably happy to leave the loose end loose. Modern "sunspot theorists" have tightened up the argument considerably, in ways that Keynes might have found congenial.

Lucas, of course, changes the answer to yes. Was this change motivated by empirical evidence that subjective expectations match the conditional expectations generated by models—or even that actual expectations are unbiased and efficient? No. Indeed, Edward Prescott has boldly asserted that "surveys cannot be used to test the rational expectations hypothesis" (1977, p. 3). Rather, economists are supposed to convert to rational expectations (RE) because of the unloveliness of the *ad hoc* expectational mechanisms that preceded it and because RE is

*Princeton University, Princeton, NJ 08544. I am grateful for stimulating discussions or correspondence with Ben Bernanke, Andrew Caplin, Mark Gertler, Stephen Goldfeld, David Romer, Andrei Shleifer, Robert Solow, and Lawrence Summers. A version of this paper with footnotes and complete references is available on request to the author.

more consistent with their (unverified) worldview that people always optimize at all margins. As Thomas Sargent put it: "Research in rational expectations...has a momentum of its own...that...stems from the logical structure of rational expectations as a modeling strategy" (1982, p. 382). The momentum, you will note, does not stem from empirics. I leave it to you to decide whether these criteria are more like those that led physicists to dump Newton in favor of Einstein, or those that led artists to abandon Manet to follow Picasso.

Modern Keynesians are split on this question. To some, the theoretical appeal of RE and the general idea that expectations should respond to policy changes are sufficient reason to conclude that "rational expectations is the right initial hypothesis." Others harbor doubts. I think the weight of the evidence—both from directly observed expectations and from indirect statistical tests of rationality (usually in conjunction with some other hypothesis)—is overwhelmingly against the RE hypothesis. Furthermore, RE is not without theoretical difficulties. We all know that RE models often have multiple equilibria. More fundamentally, RE is theoretically coherent only in the context of a single agreed-upon model. In an economy in which different people hold different views of the world, the very notion lacks clarity. For example, if Paul Volcker announces today that on New Year's Day he will raise $M1$ by 20 percent, I imagine Lucas and I will make different revisions in our expectations for, say, real GNP in 1987. Whose expectations are "rational?" Heterogeneous beliefs pose serious theoretical problems for RE. As scientists, then, I think we should be hesitant to embrace RE.

II. Is there Involuntary Unemployment?

Keynes said, nay screamed, yes. Lucas not only says no, but questions whether the phrase has meaning. In his words, "To explain why people allocate time to...unemployment we need to known why they prefer it to all other activities" (1986, p.38). Notice the words *allocate* and *prefer*. In his view, the unemployed are engaged in intelligent search or purposeful intertemporal substitution. He scoffs at the Keynesian tradition which, "by dogmatically insisting that unemployment be classified as 'involuntary'...simply cut itself off from serious thinking about the actual options unemployed people are faced with" (1986, p. 47).

This is a tough question to adjudicate on scientific grounds since the issue is largely definitional and, as Lewis Carroll pointed out, everyone is entitled to his own definitions. In Lucas's view, a person laid off from a job can, presumably, shine shoes in a railroad station or sell apples on a street corner. If he is not doing any of these things, he must be *choosing* not to do so. Both statements like this and reactions to them tend to be polemical. I guess dogmatism is in the ear of the beholder.

However, a few pertinent facts should leaven the ideological debate. First, when the unemployment rate rises, it is layoffs, not quits, that are rising while consumption falls rather than rises—all of which are bad news for search theory. Second, real wage movements are close to a random walk—which is bad news for the intertemporal substitution approach. Third, unemployment is heavily concentrated among the long-term unemployed; in 1985, for example, people who were jobless for 27 weeks or more constituted 54 percent of all unemployment and the expected duration of a complete spell of unemployment was 31 weeks. Can that be intertemporal substitution? Fourth, unemployed workers normally accept their first job offer, and those who are looking for work spend an average of only 4 hours per week on search activity. That hardly suggests a predominant role for search in explaining unemployment.

III. Do Wage Movements Quickly Clear the Labor Market?

Keynes certainly thought not, for such reasons as trade union aggressiveness, custom and inertia, and outright stubbornness. Lucas answers yes—though perhaps only in a broad sense. He has, for example, cited approvingly the competitive contract equilibrium approach in which workers have 100 percent unemployment insurance and, because of indivisibilities, are chosen randomly to work either, say, 40 hours a week or zero—meaning, of course, that unemployed workers have higher utility than employed ones. In Lucas's opinion, there is "no reason to believe" that competitive models of labor markets that treat unemployment like leisure commit "a serious strategic error."

No reason? I think the preponderance of the evidence says otherwise. Unemployment insurance replaces only about 40 percent of lost earnings. Lately, only about one-third of the unemployed collect it. Where is the evidence that the unemployed are happier than the employed? Most economists think Lucas's distinguished predecessor at the University of Chicago had it right when he wrote, "Under any conceivable institutional arrangements, and certainly those that now prevail in the United States, there is only a limited amount of flexibility in prices and wages." And it is hard, for me at least, to look at what has gone on in this country—not to mention in Europe—since 1974 and see clearing labor markets. That the market-clearing approach caught on in this environment is testimony to Lucas's keen intellect and profound influence, not to economists' respect for facts.

More than just casual empiricism supports this view; numerous formal econometric studies reject the market-clearing hypothesis against some sort of disequilibrium alternative. Unfortunately, it is usually spot-market clearing that is rejected. Equilibrium contracting models in which the wage plays little or no short-run allocative role are difficult to formulate econometrically, much less to reject. Indeed, it is hard to know what observations could contradict such models; theory just leaves too many open possibilities.

Nonetheless, certain observations are worth making. For one, several authors have pointed to interindustry wage differentials that are persistent across both time and space—differentials which are not easily squared with market clearing. Theoretically, we know that the wage rate may not be able to clear the labor market in a world of imperfect information—not even in the long run. Of course, that efficiency wage models can be built does not imply that they describe reality. But it does mean that market-clearing models have no particular claim to the theoretical high ground.

In sum, the scientific basis for modeling labor markets—or goods markets for that matter—as continuously clearing escapes me.

IV. Is the Natural Rate of Unemployment a Strong Attractor for the Actual Rate of Unemployment?

Keynes thought not. Indeed, in his revolutionary zeal, Keynes spoke loosely (loose talk was a problem for Keynes) of an "unemployment equilibrium"—which would seem to deny the natural rate any attractive force at all. Lucas answers in the affirmative.

Modern Keynesians have long had trouble with the master's notion that the economy could equilibrate below full employment; they prefer to think of unemployment as a long-lasting disequilibrium. In the United States at least, the validity of the natural rate hypothesis has not been at issue for a long time. The argument, instead, is over whether the speed of convergence to the natural rate is rapid or glacial.

On this, the American evidence is unequivocal and the European evidence is overwhelming. The U.S. civilian unemployment rate peaked at 8.9 percent in May 1975 and then took almost three years to get back down to 6 percent. It then peaked again at 10.7 percent in November-December 1982; now, four years later, it has yet to fall below 6.7 percent for even a single month. Some will argue that 7 percent is now the natural rate, without worrying much about how it grew so high. My view is that a theory that allows the natural rate to trundle along after the actual rate is not a natural rate theory at all.

In Europe, the evidence is far more compelling. Unemployment rates rose more or less steadily from 1974 to 1985—from 3 to over 13 percent in Britain, from 2.8 to 10.5 percent in France, and from 1.6 to 8 percent in Germany. Some young men in these countries have *never* held a job and may never be productive workers. Facts like these have prompted several authors to seek models which explicitly reject the natural rate hypothesis in favor of hysteresis. And recent econometric work suggests hysteresis in postwar U.S. real GNP as well. It may well be that Keynesians caved in too readily to the natural rate hypothesis.

V. Is there a Reliable Short-Run Philips Curve?

Keynes, of course, did not answer this question; the Phillips curve came later. I include it on the exam because Lucas and Sargent made it central to their attack on Keynesian economics. The alleged failure of the Phillips curve was their main piece of evidence that empirical Keynesian models "were wildly incorrect, and that the doctrine on which they were based is fundamentally flawed." (Please notice the adverbs.)

This charge was repeated so often and with such certitude that it became part of the conventional wisdom. Unfortunately, it is, to coin a phrase, wildly incorrect. The fact is that, the Lucas critique notwithstanding, the Phillips curve, once modified to allow for supply shocks (any one of several variables will do), has been one of the best-behaved empirical regularities in macroeconomics—much better behaved, in fact, than we had any right to expect. A long list of studies supports this conclusion. Nonetheless, Lucas continues to speak of the Phillips curve as an econometric basket case.

Let me anticipate the obvious objection that saving the Phillips curve after the fact by adding a supply variable does not absolve it of its *ex ante* forecasting errors. It is true that, while Robert Gordon's latest Phillips curves fit the data well, his pre-1972 equations do less well. And they did not predict the rise and fall of OPEC. But there is no sense in which new classical models either anticipated the error or pointed to the solution; like Keynesian models, they were designed to analyze demand shocks. Events in the 1970's and 1980's demonstrated to Keynesian and new Classical economists alike that Marshall's celebrated scissors also comes in a giant economy size. It is a debater's tactic, and a poor one at that, to claim that supply shocks are outside the purview of Keynesian economics.

VI. Does a Change in the Money Supply have Real Effects?

Keynes and the Keynesians answered yes, without bothering to distinguish between anticipated and unanticipated changes. Lucas and the Lucasians answer that money has real effects only if it is misperceived. In their view, a properly perceived injection of money is like a currency reform.

Here, again, the weight of the econometric evidence (though certainly not all of it) suggests that Keynes had the right answer after all. Robert Barro's alleged empirical demonstration that only unanticipated money has real effects did not hold up. Perceived changes in money are not neutral.

VII. Does Social Welfare Rise when Business Cycles are Limited?

Keynes tacitly, but unequivocally, answered yes. If asked for proof, he probably would have chuckled with the condescension of the British upper crust—which is hardly a scientific attitude.

Lucas is carefully agnostic, but clearly leans toward the answer no. He has long been sympathetic to the idea that successful stabilization policies that smooth business cycles may actually decrease welfare. Prescott is less circumspect. Without bothering to draw any distinction between modeling a conclusion and proving it, he asserts that "costly efforts at stabilization are likely to be counterproductive" because "economic fluctuations are optimal responses to uncertainty in the rate of technological change." Clearly, Harberger triangles look bigger and Okun gaps smaller near lakes than near oceans. Is Prescott's attitude more scientific than Keynes'?

I think it is worth taking a moment to explain why Lucas believes that the potential gains from stabilization policy are so small. The postwar standard deviation of log quarterly consumption around trend is about .013. Lucas asks an infinitely lived consumer living under perfect capital markets how much he would be willing to give up to reduce this small standard deviation to zero. Unsurprisingly, the answer comes back: not much. So Lucas concludes that "the post-war business cycle is just not a very important problem in terms of individual welfare." That is a stunning assertion, especially when juxtaposed against the conventional wisdom that governments rise and fall on the vicissitudes of the business cycle.

Lucas's conclusion, it seems to me, ignores a few pertinent facts. First, the cycle is not mainly in consumer expenditures, much less in consumption. Indeed, there is virtually no cycle at all in spending on nondurables and services. Are large swings in consumer durables, in inventories, and in fixed investment all socially costless? Don't these ups and downs impose serious adjustment costs and dislocations on society?

Second, Lucas's calculation assumes that cyclical fluctuations take place around an unchanged trend, with booms as likely as recessions. But what if recessions leave permanent scars on either labor or capital or productivity? What if there is hysteresis, so the natural rate hypothesis fails? What if there is a systematic tendency for output to be too low on average? Then the Keynesian goal of filling in troughs without shaving off peaks starts to make sense.

Third, Lucas ignores a variety of psychological, sociological, and physiological costs which many people feel are important. Against Lucas's benign view of the cycle compare the opinion of Martin Luther King, who wrote that "In our society, it is murder, psychologically, to deprive a man of a job or an income. You are in substance saying to that man that he has no right to exist." The truth, I think, lies somewhere between Lucas and King.

Finally, it is important to remember that cyclical losses are not distributed uniformly, as Lucas assumes; instead, most people lose little while a minority suffers much. Let me illustrate with some simple calculations. Suppose everyone has log utility and consumes \$3000 per quarter. Let a severe recession reduce consumption 4 percent. Utility falls 4.1 percent, which is no big deal, especially since every down is matched by a subsequent up. This is Lucas's world.

Now change utility to the Stone-Geary form: $U = \log(C - \$1500)$. Here a 4 percent drop in consumption reduces utility by 8.3 percent. That seems a bigger deal. Finally, let the cycle instead reduce the consumption of 10 percent of the population by 40 percent while the other 90 percent loses nothing. (Note that I am allowing very generous unemployment insurance here.) With the Stone-Geary utility function, mean utility declines 16.1 percent. Now we're talking real utils.

Lucas will, of course, counter that any such problem is best dealt with by better unemployment insurance, not by stabilization policies that interfere with free-market allocations. The same logic says that fire and theft insurance—where moral hazard problems are certainly less severe—obviate the need for fire and police departments. Isn't prevention better than insurance?

However, Lucas's challenge to the Keynesian presumption that smaller cycles are better cycles needs to be addressed scientifically. And, since we can't observe cyclical fluctuations in utility, that requires the use of theory. The relevant theory is, I think, just beginning to be developed in the burgeoning literature on monopolistic competition and aggregate demand externalities. It would be foolish to say that a definitive answer is in hand; but some good answers may be on the horizon.

VIII. Must Macroeconomics be Built Up from Neoclassical First Principles?

Keynes answered no. A practical man living in a complex world, he would not close his eyes to apparent deviations from narrow-minded concepts of optimizing behavior —nor even to gross deviations from rationality. He believed in modeling behavior as it was. Witness his defense of money illusion in labor supply:

> Now ordinary experience tells us...that a situation where labor stipulates...for a money-wage rather than a real wage, so far from being a mere possibility, is

> the normal case.... It is sometimes said that it would be illogical for labour to resist a reduction of money-wages but not to resist a reduction of real wages.... But, whether logical or illogical, experience shows that this is how labour in fact behaves. [p. 9]

Lucas and other new classicists take a different view. They emphasize the importance of building up macroeconomic relationships from sound microfoundations, by which they mean the solutions to dynamic, stochastic games. Lapses from what Lucas called "the only 'engine for the discovery of truth'" are one of the chief grounds on which Keynesianism is branded unscientific.

Now, neither side is hostile either to first principles or to factual accuracy. We all agree that the ideal macro theory would be built up logically from first principles and would explain the data well. But we also agree that such a theory is a long way off. The issue is how religiously we must adhere to frictionless neoclassical optimizing principles until that glorious day arrives. Here the devoutness of American economists distinguishes us from our colleagues in other lands. But which attitude leads to better science? Is it better to start deductively from axioms or inductively from facts? When the time comes to choose between internal consistency and consistency with observations, which side should we take? Must we be restricted to microfoundations that preclude the colossal market failures that created macroeconomics as a subdiscipline?

Here followers of Keynes and followers of Lucas often part company. Like Keynes, modern Keynesians are inclined to begin by "taking things as they are"; rigorous optimizing explanations for what they observe (such as nominal wage contracts) can come later. The important thing is to make sure our models are congruent with the facts. Lucasians, it seems to me, reverse the sequence. They want to begin with fully articulated, tractable models and worry later about realism and descriptive accuracy.

This is a judgment call; but I judge the Keynesian approach more scientific. First, good science need not always be built up from solid microfoundations. Thermodynamics and chemistry, for example, have done pretty well without much micro theory. Boyle's Law applies directly to aggregates, much like the marginal propensity to consume. And the microfoundations of medicine are often very poor; yet much of it works. Empirical regularities that are formulated and tested directly at the macro level *do* have a place in science.

Second, it is far from clear that the particular first principles selected by new classical economists deserve to come first. Why don't people know the money supply or the price level within very small margins of error? Who imposed a cash-in-advance constraint? Why should price move to equate supply and demand in markets with asymmetric information? Why, Keynes might ask, are these postulates more acceptable as first principles than nominal wage contracting?

Third, the model of man as a strongly rational maximizer is not the only option open to theorists. There are theories of "bounded rationality" and of "near rationality." Even within the strict optimizing framework, neoclassical tangencies are not the only, nor even the most likely, alternative. Pervasive lumpy transactions costs lead to "the optimality of usually doing nothing," meaning that it rarely pays to change your decision variable, even if it is not set at the frictionless "optimal" value. In a word, near rationality is full rationality. It is continuous optimization that would be irrational.

Direct empirical evidence on individual behavior is difficult—some would say impossible—to come by. But what little we know from experiments by psychologists like Daniel Kahneman and Amos Tversky and others does not suggest that *homo sapiens* behaves like *homo economicus*. (Perhaps that is why they have different names.) Inconsistent choices are common. People put too much weight on what has happened to them and to their friends and too little on statistical evidence. Framing of the question matters. The von Neumann-Morgenstern axioms are routinely violated. It is remarkable how little impact this evidence has had on modern economics. Is that scientific detachment or religious zealotry?

So I have come to the end of my exam with the conclusions you might have guessed at the outset: that when Lucas changed the answers given by Keynes, he was mostly turning better answers into worse ones; that modern Keynesian economics, though far from flawless, has a better claim to being "scientific" than does new classical economics.

REFERENCES

Friedman, Milton, "The Role of Monetary Policy," *American Economic Review*, March 1968, *78*, 1–17.

Keynes, J. M., *The General Theory of Employment, Interest, and Money*, London: Harcourt, Brace and World, 1936.

Lucas, Robert E., Jr., "Models of Business Cycles," paper prepared for the Yrjo Jahnsson Lectures, Helsinki, Finland, mimeo., March 1986.

Prescott, Edward, "Should Control Theory be Used for Economic Stabilization?," in Karl Brunner and Alan Meltzer, eds., *Optimal Policies, Control Theory, and Technological Exports*, Vol. 7, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 13–38.

______, "Theory Ahead of Business Cycle Measurement," Federal Reserve Bank of Minneapolis Research Department Staff Report 102, February 1986.

Sargent, Thomas J., "Beyond Demand and Supply Curves in Macroeconomics," *American Economic Review Proceedings*, May 1982, *72*, 382–89.

Rational Models of Irrational Behavior

By GEORGE A. AKERLOF AND JANET L. YELLEN*

The General Theory revolutionized macroeconomics with its assertion that classical economics contained a fundamental error. Keynes argued that a capitalist economy could possess equilibria characterized by persistent involuntary unemployment. He also showed that aggregate demand would play a crucial role in determining output and employment. Keynes' analysis accorded with common sense and casual observation fifty years ago. In our opinion, it still does so. Why then is there a crisis in Keynesian economics?

Keynesian analysis violates the commonly regarded sine qua non of good economic theory—a microeconomic foundation based on perfectly rational, maximizing behavior. In our reading, economists have accorded the assumption of rational, self-interested behavior unwarranted ritual purity, while alternative assumptions—that agents follow rules of thumb, that psychological or sociological considerations matter, or that, heaven forbid, they act downright irrationally at times—have been accorded corresponding ritual impurity. This association between "impure" assumptions and ritual pollution has had the ill effect of confusing the esthetic task of economics—which is to provide clear logic for analyzing economic phenomena—with the agenda of economics—which is to explain the economic events of the real world. Keynesian theory, with its partial reliance on psychological, sociological, and rule-of-thumb behavior to derive departures from full employment and Pareto optimality, is the worst casualty of this failure to dissociate esthetic from agenda. If agents really behave according to impure assumptions, is it not likely that the best models to fulfill the agenda will mirror that behavior?

According to standard Keynesian analysis, the key departure from self-interested, maximizing behavior is the assumed stickiness of money wages. Keynes argued that because of the importance to workers of relative wages, they quite typically resist money wage reductions in circumstances where they are willing to assent to real wage reductions. He recognized that "it is sometimes said that it would be illogical for labour to resist a reduction of money-wages but not to resist a reduction of real wages." But, he concluded, "whether logical or illogical, experience shows that this is how labour in fact behaves" (1935, p. 9).

Keynes' willingness to build a theory which was a pastiche of optimizing behavior and sociological/psychological rule-of-thumb behavior grounded in the observation of how people appear to behave was probably due to his conviction that the central features of his theory in no way hinged on the seemingly illogical behavior of workers. *The General Theory* makes clear that money wage stickiness is *not* in Keynes' opinion the ultimate cause of involuntary unemployment; indeed, due to the adverse effects of falling prices on demand, involuntary unemployment might possibly be more severe in its absence. But most Keynesians accept the verdict of Patinkin and Pigou that, in the presence of a real balance effect, the aggregate demand curve is *not* vertical and thus a full-employment equilibrium exists. What to Keynes was a minor assumption in a theory rationalizing business cycles is now interpreted as the key assumption.

During the past two decades Keynesian theorists have struggled to formulate a "sensible" microeconomic foundation for Keynesian economics based on individualistic optimizing behavior, by relaxing the assumptions of the perfectly competitive

*University of California, Berkeley, CA 94720. We are grateful to the Sloan Foundation, Guggenheim Foundation, and the National Science Foundation for support under grant SES86-005023, and to Daniel Kahneman and Jim Wilcox for helpful comments.

Walrasian model and introducing instead a dizzying array of market imperfections: asymmetric information, incomplete contingent claims markets, staggered contracts, transactions costs, imperfect competition, specific human capital, efficiency wages, etc. Have these efforts been successful? Not entirely. While each of these innovations has enriched economics by modeling important aspects of reality, the introduction of these imperfections has still not provided a *total* rationalization of Keynesian economics when judged according to the rule that the proposed theory be *fully* consistent with rational optimizing behavior and the absence of any unexploited gains from trade. In the end, it invariably turns out either that there is an unrealistic assumption or that some clever, complicated neoclassical contract will eliminate involuntary unemployment. For example, most versions of efficiency wage theory which are grounded in optimizing behavior suffer from the "defect" that there exist contracts which, while rarely observed, are feasible in principle (for example, employment bonds, job auctions, tournament contracts) and can eliminate involuntary unemployment if firms can establish reputations for trustworthy labor relations.

This paper begins with the premise that theory which fits the real world will be based on assumptions that individuals are not fully rational. It would be simply unscientific to proceed otherwise. For indeed, individuals may actually suffer from money illusion, follow rules of thumb, or give weight to considerations of fairness and equity in economic matters. Section II reviews some evidence which indicates that individuals do behave in such ways, and, furthermore, that the sticky money wages assumed in Keynesian models are consistent with the known irrationalities of human behavior. A second justification for old-style Keynesian models without full rationality, is that the departures from rationality needed to generate Keynesian business cycles are not very great. Section I argues that rejection of Keynesian business cycles on the grounds that fully rational models must be money neutral is based on a faulty implicit assumption—that significant deviations from rationality are required to rationalize the Keynesian model.

I. How Irrational is Keynesian Economics?

Keynes argued that nominal shocks to aggregate demand (due, say, to changes in the money supply)—whether anticipated or unanticipated—can cause prolonged departures of the economy from full employment. Such departures simply cannot occur in any model that assumes *fully* rational optimizing behavior, including rationally formed expectations, unless nominal prices exhibit inertia. The logic is simple. Rational agents should only care about real magnitudes. If so, any optimizing model with a unique equilibrium will be "money neutral." A cut in the supply of money should just cause a proportional reduction in prices and wages with no change in employment, output, or real wages. This conclusion in no way depends on the assumption that markets are perfectly competitive, or even that markets clear.

Demand-generated business cycles require nominal price rigidity. Such rigidity could be explained by "menu costs" associated with nominal price changes, or by money illusion. In the case of menu costs, one can hardly imagine that the objective costs of price changes are sufficiently large to explain business cycles. The leading argument against money illusion as a factor generating business cycles is the implausibility that agents persistently pass up opportunities for gain. But how large must transactions costs be? Or how foolish are agents whose behavior exhibits money illusion? Are individuals who change wages and/or prices inertially leaving the proverbial $500 bills on the sidewalk? Or are they failing to stoop to pick up a few pennies?

Near Rational Behavior. According to previous work by ourselves (1985a,b; 1986) and others (N. Gregory Mankiw, 1985; Olivier Blanchard and Nobuhiro Kiyotaki, 1985), many forms of seemingly irrational behavior may really be "near-rational." By this, we mean that agents have relatively wide latitude for deviating from full optimization without incurring significant losses. In mathematical terms, this is a consequence of the envelope theorem which states, in effect, that the impact of an exogenous shock on a fully maximizing agent is identical, up to a first-order of approximation, whether he optimally changes his decision variable in response to a shock, or instead responds inertially. Stated differently, inertial, or rule-of-thumb behavior typically imposes losses on its practitioners, relative to the rewards from optimizing, which are second-order. Thus, slight relaxation of the standards for "good" model building—so as to tolerate behavioral assumptions entailing suitably small losses from nonmaximizing—significantly enlarges the range of behavior to be considered. For example, it turns out that in many contexts, the inertial adjustment of nominal wages and prices is near rational. Staggered nominal contracts in the style of John Taylor (1979), which do not quite optimally make use of newly available information because they keep nominal prices constant for two periods, turn out to be near-rational as also are the stock adjustment responses assumed to characterize money demand, consumption and investment in many Keynesian models. It might be thought that near-rational theories must be close to fully rational theories; but this intuition is in fact incorrect.

An Example. The logic of the difference between near rationality and full rationality can be explained in the context of a simple example. According to this example, if firms are monopolistically competitive, and a fraction of them change prices inertially in response to money supply changes, then significant business cycles result. However, nonmaximizing firms fare only insignificantly worse than optimizing firms. Consider an economy with identical, monopolistically competitive firms selling differentiated products. To make things simple, imagine that output can be produced costlessly. Further, assume that each firm's sales depends on its price relative to its rivals and on aggregate demand which, again for simplicity, is proportional to real balances—the nominal supply of money divided by the average price level. Finally, imagine that each firm sets its price in Bertrand fashion, choosing the price that maximizes profits, taking rivals' prices as given, and the firms have settled into a Bertrand equilibrium. If all firms are fully rational and the Fed cuts the money supply, the new equilibrium will be exactly like the old in real terms. All that happens is that each firm's nominal price falls in proportion to the money supply cut. Money is neutral.

Now consider what would happen if some proportion of firms are nonmaximizers who follow an inertial rule in altering prices. To take an extreme case, suppose the nonmaximizers leave their price unchanged following the cut in the money supply. How much do they lose? The answer is that their losses are second-order. If the money supply fell by a percentage ε, the inertial firms will make a pricing error which is proportional to ε, but incur losses which are proportional to ε^2—second-order in terms of the shock to the system. However, the decline in real balances (and hence output) resulting from the cut in M is first-order—proportional to ε. How can this be? Recall that under normal circumstances, any optimal decision is made by just balancing the marginal gains and losses from a change in the decision variable. The price-setting competitors optimally set prices by balancing the marginal gain of greater sales from a reduction in the price against the loss due to lower profit on each unit sold. An optimum has not been reached until the firm is indifferent, to a first-order of approximation, about its price. The firm's profit, as a function of its own price, is almost flat in the neighborhood of an optimum. Accordingly, the firm has latitude for making a relatively large error, without suffering large losses. An error of size δ causes a loss proportional to δ^2. In this context, inertial price setting is almost costless even though its macroeconomic impact is significant.

The logic of the preceding example concerning prices suggests a similar defense of the standard Keynesian assumption of nominal wage rigidity. Nominal wage rigidity is near-rational if the firms' profits are a continuously differentiable function of the wage they pay. In the efficiency wage paradigm, for example, the "productivity" of workers, broadly defined, is an increasing function of the real wage they receive. Firms optimally set wages by equating the marginal gains from lower wage cost per worker with the losses from lower productivity per worker; accordingly, inertial wage setting is near-rational. A firm that (nonoptimally) fails to cut wages in a recession gets some reward from its behavior—higher morale, lower turnover, etc. And while the behavior may not be strictly optimal, it is almost optimal since, for a firm which was initially optimizing, the rewards and costs of wage cuts were exactly balanced to start with.

II. How Rational Are Economic Agents?

We argued above that nonrational elements should be brought into macro models, and that, in many models, not much irrationality is needed to produce business cycles. In this section we shall argue that psychology and sociology provide natural explanations for sticky money wages. The New Classical Economics' conclusion, that nominal variables are proportional to the expected money supply, is a singularity, in no way predicted by the judgmental errors and concern with equity which is well documented in psychology and sociology.

Cognitive Biases. Twenty years ago it was widely believed that in most cognitive judgments people acted as intuitive scientists. However, two decades of work by cognitive and social psychologists have unearthed a large variety of ways in which individuals' judgments exhibit systematic errors relative to the scientific, objectively rational model.

People use at least three heuristics which generate biases in their decisions. According to the *availability heuristic*, they depend more than they should on "salient" information which is easily retrievable from memory. According to the *representativeness heuristic*, they act as if stereotypes are more common than they actually are, and in *anchoring*, they let their judgments be overly reliant on some initial "anchoring" values. Max Bazerman (1986) has enumerated thirteen distinct judgmental mistakes that are due to the use of the three heuristics. These judgment errors are too numerous and too frequently documented in the laboratory with salient examples in the field to be easily dismissed as unimportant.

The question arises whether there are cognitive biases that suggest potential reasons for money wage stickiness. The most natural explanation of sticky money wages stems from anchoring. In a typical anchoring experiment, one finds that "irrelevant" initial conditions affect outcomes. For example, in a classic anchoring experiment, Daniel Kahneman and Amos Tversky (see Bazerman) spun a roulette wheel, and then asked subjects to estimate the number of African states with representatives in the United Nations—using the number obtained by spinning the roulette wheel as the initial estimate. For those whose initial estimate from the roulette wheel was 10, the median estimate was 25; for those whose initial estimate from the roulette wheel was 65, the median estimate was 45!

Could anchoring explain sluggish adjustment of money wages? It certainly could if last-period's money wage acts as an anchor which influences this period's wage settlement. As we shall see presently, in the discussion of *fairness*, people's views of fair money wages apparently are anchored in the current *money* wage.

Fairness. It is hard to believe that payments are not jointly determined by market forces and fairness. Steven Allen, Robert Clark, and Daniel Sumner (1984) have pointed out that some firms have voluntarily added some indexation to benefits paid to already retired employees. While these payments undoubtedly had some beneficial effect in enhancing the firms' reputations with current and prospective employees, a more natural explanation for these payments is not pure profit maximization by the firms, but a commitment to fair behavior.

In order for fairness to play a role in determining contract outcomes there must be a discrepancy between fair and market clearing outcomes. Kahneman, Jack Knetsch, and Richard Thaler (1986) have provided indisputable evidence of such a divergence in a recent interview study. In this study they described a variety of situations to randomly sampled telephone interviewees, and then asked whether the market-clearing solution was fair. For example, they asked whether it would be fair for a hardware store to charge more for snow shovels following a snowstorm: 82 percent think it unfair and 18 percent think it fair. As a second example, they told interviewees that a small photocopying shop has one employee earning $9 per hour. Business continues to be satisfactory but a factory in the area has closed and unemployment has increased. Would it be fair to reduce the wage paid to $7 an hour now paid elsewhere to newly hired workers with similar talents? Eighty-three percent think it unfair.

Anchoring played a crucial role in most of the interview results in this study. Most questions involved the fair response to a change from some initial situation. It was, in general, considered unfair for one party to benefit relative to the initial anchoring situation while the second party lost. Thus in the hardware store example, it was unfair for the store to profit while customers paid more for snow shovels; similarly, it was unfair for the firm to cut the wages of its existing employees.

Importantly for Keynesian economics the Kahneman-Knetsch-Thaler experiments indicated the presence of anchoring based on current *money* wages. Respondents felt it fair for a company whose business was bad to raise wages by only 7 percent when there was 12 percent inflation, but unfair to cut wages by 5 percent if there was no inflation.

These findings are consistent with the standard sociological theory of fairness known as equity theory. (For an excellent survey see Roger Brown, 1986.) Equity theory (due to J. Stacy Adams and George Homans) predicts that in exchanges, "outcomes" (rewards) relative to the "investments" must be equal between the parties to the transaction. While this theory is specifically derivative from economics, there is an important difference between its predictions and those of standard economic theory. In equity theory, the investments and outcomes are *subjectively* measured. The provider of labor services may *subjectively* value his or her services by more than the buyer of those services. Then, if we take the gap between the ratio of "reward" relative to investment on the part of the buyer and the seller, we have a prediction of the extent of resentment that parties in an exchange will experience. Because market-clearing contracts may be viewed as inequitable, the attempt to impose market-clearing terms on a transaction can easily generate resentment.

The existence of resentment does not, of course, imply that economic outcomes cannot occur which are resented. But in labor contracts there are good reasons why employers will wish to temper the resentments of their employees and accordingly may offer contracts that are not market clearing. Quite simply, resentment caused by a sense of unfair treatment is likely to translate into poor performance by workers who can exercise discretion in the performance of their work. Even if supervision and monitoring are feasible at low cost, it may not pay firms to monitor their employees too closely. A recent study by Edward Deci, James Connell, and Richard Ryan (1985) has shown that workers who are given more detailed rules and are more closely monitored experience less job satisfaction, are less motivated, and place more importance on such external rewards as compensation; in contrast, those who are less controlled achieve greater satisfaction in mastering their jobs. If firms are to take advantage of the self-motivation that comes with on-the-job autonomy, then they must ensure that workers perceive themselves to be fairly treated. The cost can easily involve paying wages in excess of market clearing. Accordingly, we would expect *perceptions* of fairness to play a role in determining wage contracts and anchoring to cause money wages to be sticky.

III. Conclusion

The bad press that Keynesian theory has recently received from maximizing, super-rational theory is simply undeserved. The assumptions required to motivate Keynesian economics are quite consistent with the behavioral regularities documented by psychologists and sociologists. This motivation is in no way tortured out of complicated assumptions and models. It is highly natural. Keynesianism, both as theory and explanation of the facts, is alive and well on its fiftieth birthday. Happy Birthday, *General Theory*!

REFERENCES

Akerlof, George and Yellen, Janet, (1985a) "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly Journal of Economics*, Supplement 1985, *100*, 823–38.

______ **and** ______, (1985b) "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *American Economic Review*, September 1985, *75*, 708–20.

______ **and** ______, "The Macroeconomic Applications of a Dynamic Envelope Theorem," mimeo., University of California-Berkeley, 1986.

Allen, Steven G., Clark, Robert L. and Sumner, Daniel A., "Post-Retirement Adjustments of Pension Benefits," NBER Working Paper No. 1364, June 1984.

Bazerman, Max, *Judgment in Managerial Decisionmaking*, New York: Wiley & Sons, 1986.

Blanchard, Olivier and Kiyotaki, Nobuhiro, "Monopolistic Competition, Aggregate Demand Externalities and Real Effects of Nominal Money," NBER Working Paper No. 1770, December 1985.

Brown, Roger, *Social Psychology, the Second Edition*, New York: Free Press, 1986.

Deci, Edward L., Connell, James P. and Ryan, Richard M., "Self-Determination in a Work Organization," mimeo, University of Rochester, 1985.

Kahneman, Daniel, Knetsch, Jack and Thaler, Richard, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, *76*, 728–41.

Keynes, J. M., *The General Theory of Employment, Interest and Money*, London: Harcourt, Brace and World, 1935.

Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly," *Quarterly Journal of Economics*, May 1985, *100*, 529–37.

Taylor, John, "Staggered Wage Setting in a Macro Model," *American Economic Review Proceedings*, May 1979, *69*, 108–13.

Residual Differences by Sex: Perspectives on the Gender Gap in Earnings

By CLAUDIA GOLDIN AND SOLOMON POLACHEK*

Relative to men, women receive lower wages and are employed in lower paid and more menial occupations. Because these unequal economic outcomes exist for a majority group, widely dispersed in society, the subject of gender differentials is central to understanding discrimination. Most measures of sex discrimination assess whether the market differentially rewards otherwise similar men and women. Differences in earnings generated by differences in "female" and "male" characteristics represent the "explained" portion of the gender gap, while differences in earnings generated by differences in "rewards" per attribute are termed an "unexplained residual." It is the unexplained portion that has been identified with market discrimination, despite many potential biases. If discriminatory market signals discourage investments, and if socialization, not individual choice, determines outcomes, the residual may be a lower bound to the true value of discrimination. Discrimination, however, would be overestimated if differential investments are motivated by individual choice based on optimal wealth-maximizing behavior, such as the division of labor within the home.

The consensus figure for sex discrimination today is about 50 to 60 percent of the earnings gap for full-time male and female workers when actual experience is included but occupational categories are not (recent summaries include Janice Madden, 1985; Francine Blau and Marianne Ferber, 1987). Thus an initial gap of about 40 percentage points is reduced to about 23 percentage points. Refinements to the data on college education and job characteristics suggest further reductions (T. N. Daymont and P. J. Andrisani, 1984; Randall Filer, 1983).

Because all readily available data sources are for the post-1960 period, concern has centered on the recent past. The earnings gap declined in the early 1950's, stayed virtually constant from the mid-1950's to 1980, and narrowed again after 1980 (Blau and Andrea Beller, 1986, however, find evidence of a closing from 1971). Movements in the residual over the brief period since 1960 are difficult to distinguish given the wide range of results and data sets. Over a far longer period of time, however, considerably more can be discerned. Data spanning the last 165 years indicate that the ratio of female to male earnings, adjusted for time inputs, increased from about 0.35 in 1820 for manufacturing workers, to 0.45 in 1890 for all workers, and to over 0.65 in 1985 (Goldin, 1986), with most of the increase for all workers occurring from 1890 to 1930.

Disaggregated data for various periods from 1890 to 1940 indicate that a far greater percentage of the log difference between male and female earnings can be explained in historical data than in current survey data, implying a secular increase in discrimination despite a narrowing in the wage gap. We contend that the apparent rise in discrimination is illusory, coming about because of biases in the traditional discrimination measure. We also illustrate related propositions using current cross-sectional data.

[†] *Discussants*: Francine D. Blau, University of Illinois; Sherwin Rosen, University of Chicago.

[*] Departments of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, and SUNY-Binghamton, Binghamton, NY 13901, respectively. Section II borrows heavily from Goldin, 1986. We thank Yesook Merrill for computer assistance.

We offer two perspectives on the gender gap. The first concerns the historical experience, and in it we suggest that advances in attributes need not be part of the explained portion. In the second, that on the recent period, we demonstrate that differences in rewards per attribute need not be due to labor market discrimination.

I. Decomposing the Earnings Gap: The Discrimination Measure

Discrimination is often defined as that portion of the earnings gap unexplained by individual characteristics. If mean male and female earnings are given by $y_m = g_m(x_m)$ and $y_f = g_f(x_f)$, where g_m and g_f are the male and female earnings functions and the x's and y's are mean values, then $y_{mf} = g_m(x_f)$ is the earnings of males had they female characteristics and $y_{fm} = g_f(x_m)$ is the earnings of females had they male characteristics. Alternatively, y_{mf} (and y_{fm}) can be interpreted as the mean female (male) wage had women (men) a male (female) wage structure. The amount $(y_{fm} - y_f)$ is the difference in earnings explained by differences in male and female attributes. Thus $c = [(y_{fm} - y_f)/(y_m - y_f)]$ is the portion of the wage gap explained by differences in characteristics, while $d = (1 - c)$ is the unexplained portion. By convention, d is the measure of discrimination (Ronald Oaxaca, 1973). The amount $(y_m - y_{mf})$ is also an earnings differential due to attribute differences; $c' = [(y_m - y_{mf})/(y_m - y_f)]$ is also the portion of the wage gap explained by attribute differences and $d' = (1 - c')$ is another measure of discrimination. The two approaches differ only in the use of a male or female comparison earnings function. Because $x_m > x_f$ and $g'_m > g'_f$, generally, the first approach yields a larger estimate of discrimination.

Discrimination is, therefore, identified with differences in the regression coefficients between g_m and g_f (including the intercepts). But should structural earnings differences be considered discriminatory when differences in attributes are not? Attribute differences are often considered nondiscriminatory because many individual characteristics are determined well before work commences, and aspects of an individual's work history are determined prior to the current job. But society may discriminate in access to skills and comparable schooling. Further, unmeasurable differences may be induced by way of feedback effects. Gender differences in upbringing may be caused by stereotypes and norms, and societal discrimination and employee expectation of future labor market discrimination can induce differences in factor endowments.

While the process of attribute acquisition may be discriminatory, differences in the rewards to attributes need not be. The division of labor by men and women in home and work activities, for example, implies differential specialization and by implication a differential impact on earnings of identically measured attributes. Being married raises a man's wage because of increased human capital investment associated with a stronger lifetime work commitment. But being married may lower a woman's wage because of smaller investment incentives. In this case, $g'_m > g'_f$, and a marital status adjustment in a wage regression would produce misleading results. The conventional estimate of discrimination, however, assumes that different attribute rewards that are the result of family optimization, and that may occur in the absence of labor market discrimination, indicate discrimination.

II. Historical Perspectives on the Earnings Gap and Discrimination

The history of relative earnings of males and females can be constructed using a variety of historical and archival data sources beginning nearly two centuries ago with data from the agricultural and manufacturing sectors. Earnings ratios for the entire economy, however, can be constructed only for the last century. All earnings and wage data to be presented are for full-time workers and are consistent with those in most current data sets.

The wage of females relative to males was very low in the northeastern states prior to industrialization, but rose quickly wherever manufacturing activity spread. Around 1815, the ratio of female to male wages in agriculture and domestic activities was 0.288, but it rose to between 0.303 and 0.371 among manufacturing establishments at the inception of industrialization in 1820. By 1832, the average ratio in manufacturing was about 0.44, and it continued to rise reaching a level just below 0.50 in the northeastern states by 1850. Early industrialization, therefore, increased the wage of females relative to males by over 70 percent (from 0.288 to 0.50) and the ratio within the industrial sector expanded by 43 percent (from 0.35 to 0.50). The narrowing of the earnings gap in manufacturing across the nineteenth century, particularly from 1820 to 1850, resulted from an increased division of labor and use of machinery. Further, the role of industrialization depended on the crop; grain, not cotton, growing areas experienced the greatest increase in the earnings ratio. It should be noted, as well, that the ratio of boy to adult male earnings also increased over the period, suggesting that industrialization affected the productivity of certain groups more than it broke down traditional boundaries defining the role of women.

Manufacturing data provide nearly two centuries of information on the gender gap, but because the sector employed only one-third of all female workers across the last century, it is necessary to construct earnings data for a wider range of occupations. Full-time earnings for males and females underlie the ratios given in Table 1, Part A, for six major occupational groupings in three benchmark years: 1890, 1930, and 1970. Average, national, earnings ratios (row 7) are constructed by weighting by the occupational distributions (not given).

The ratio of female to male full-time earnings across the entire economy increased from 0.463 to 0.603, or by 30 percent, over the 1890 to 1970 period. The latter figure (0.603) is unadjusted for differences among full-time workers in average hours of work per week and increases to 0.663 when the implied earnings per hour are used. Thus the increase in the ratio of female to male earnings is between 30 to 43 percent over the 80 years, a finding that overturns the usual presumption that the earnings gap was as constant before 1950 as it was after. Furthermore, because the gap closed to about 1930, remaining virtually stable to about 1980, the narrowing from 1890 by about 1/3 extended over only a 40-year period. Increases in the ratio by occupational group were greatest in the professional and clerical categories, for

TABLE 1—FEMALE TO MALE EARNINGS RATIOS: 1890 TO 1970

A: Ratios of Female to Male Full-Time Earnings

	1890 w_f/w_m	1930 w_f/w_m	1970 w_f/w_m
1. Professional	0.263	0.385	0.710
2. Clerical	0.487	0.706	0.686
3. Sales	0.595	0.607	0.438
4. Manual	0.535	0.575	0.557
5. Service	0.530	0.598	0.558
6. Farm	0.530	0.598	0.589
7. (w_{fi}/w_{mi})[a]	0.463	0.556	0.603
8. $(w_f/w_m)_{1890}$[b]	0.463	0.489	0.455
9. $(w_f/w_m)_{1930}$[b]	0.534	0.556	0.507
10. $(w_f/w_m)_{1970}$[b]	0.571	0.610	0.603

B: Estimated and Approximated Coefficients and Means from Earnings Equations,

	Coefficients (g'_m, g'_f)		Means (x_m, x_f)	
	Male	Female	Male	Female
Experience, 1890	0.075	0.090	17.0	7.0
Experience squared, 1890	−0.0012	−0.003		
Education, 1890	0.02	0.010	7.0	5.4
Experience, 1970	0.045	0.025	20.0	14.0
Experience squared, 1970	−0.0005	0.0		
Education, 1970	0.065	0.070	12.7	12.6
Home time, 1970	0.0	−0.005	0.0	4.6

C: Decomposing the Change in the Log of Female to Male Earnings, 1890 to 1970

	Experience	Education	Home Time
1. $(y_{fm}-y_m)^1 -(y_{fm}-y_m)^0$[c]	0.065	0.134	0.0
2. $(y_f-y_{fm})^1 -(y_f-y_{fm})^0$	0.030	0.009	−0.023
3. $(y_f-y_{mf})^1 -(y_f-y_{mf})^0$	−0.199	0.117	0.0
4. $(y_{mf}-y_m)^1 -(y_{mf}-y_m)^0$	0.294	0.026	−0.023
Total = 0.215	0.095	0.143	−0.023

Notes and Sources: See Goldin (1986) with changes listed below. Part B: A quadratic term for job experience has been added. The mean experience levels for 1890 have been increased to account for sectors other than manufacturing. The means for work experience in 1970 are from Mary Corcoran and Greg Duncan (1979).

[a] Average of earnings (not given) for rows (1)–(6), weighted by the occupational distribution.

[b] Uses wages of subscripted year and occupational weights from column year.

[c] See text for notation definitions; 0 = 1890; 1 = 1970. Rows 1 and 3 are due to changes in differences in coefficients; rows 2 and 4 are due to changes in differences in characteristics.

which advances in education appear to have augmented both the relative earnings of females to males and the number of females employed (Goldin, 1984).

It is generally presumed that the distribution of occupations between men and women is the major determinant of the gender gap in earnings and that changes in the occupational distribution provide the primary way of altering relative earnings. Although there are only six occupational groupings in Table 1, Part A, occupational change over time for both men and women was substantial. But, as can be seen in rows 8 to 10, changes in the earnings between men and women within each of the groups were more important in altering the gender gap than were changes in the occupational mix of the two groups. For each row, the wage is held constant at that for the subscripted year, but the occupational distribution for males and females is allowed to vary by year. The ratio of female to male earnings increases far more going down the columns than it does going across the rows, findings that are robust to a more elaborate partitioning.

These results highlight the importance of the within-sector wage ratios and thus the role of factors like increased education, sectoral shifts in the demand for labor, and diminished discrimination in the general decrease of the gender gap. Thus we turn to a more analytical and less mechanical method of explaining the earnings gap at each point in time and over time.

Has our ability to explain the gap in earnings increased or decreased over time with its narrowing, and what factors account for its narrowing? It appears that the explanatory power of the conventional earnings equation, in terms of the percentage of the difference in the log of earnings that is explained, has decreased over time. The absolute value of the difference in the log of earnings that is unexplained has remained roughly (and mysteriously) constant over time, implying increased discrimination with the narrowing of the gap.

Several studies using late-nineteenth-century data have estimated earnings equations for males and females in manufacturing. In one, using a sample for 1892 (Barry Eichengreen, 1984) and having results consistent with other studies, the difference in the log of male and female earnings is 0.767 (a bit higher than the national average in manufacturing). Of this total, 0.466 to 0.492 can be accounted for by differences in the mean values of the independent variables, that is, 62.5 percent can be explained. The remaining 0.275 to 0.302 is due to differences in the coefficients, including the constant terms. Discrimination or the residual, therefore, accounts for 37.5 percent of the difference in the log of earnings, or 0.288 in absolute value.

Most recent studies find the residual, computed in this manner, accounts for 55 percent of the difference in the log of earnings. If the ratio of female to male earnings is 0.60 (i.e., $\log_e(1/0.60) = 0.511$), the explained portion is 0.230 and the unexplained portion, or the residual, is 0.281. Thus the value of the unexplained portion has remained virtually constant over time but increased as a percentage of the log difference in earnings with the narrowing of the earnings ratio. Furthermore, estimates of direct measures of discrimination using manufacturing data for 1970 (Jonathan Leonard, 1984; Paula Voos, 1986) find substantial differences between ratios of marginal products and ratios of wages. But cross-sectional (by city, by industry) estimates for 1890 find no differences between the ratio of the marginal products and that of wages by sex (Goldin, in progress).

What factors account for the increase in the log of the female to male earnings ratio? Table 1, Part B, gives the consensus coefficients and variable means from recent earnings function studies and those of the turn of this century. The variables that can be included are experience, education, and "home time." The estimated and approximated coefficients and means yield a total explained portion of 0.215, (Table 1, Part C). Of the total, 0.095 is due to changes in the experience variable, 0.143 is due to changes in the education variable, and the increase in home time reduces the total by 0.023. The two methods of decomposition described above are given in Part C, and, on average, changes in characteristics had a greater impact on experience, while changes in the coefficients had the greater effect on education. These findings are consistent with the notion that increases in female earnings within certain occupations—those for which returns to education were highest—served to narrow the earnings gap.

The extensive use of piece-rate wages in manufacturing around the turn of the century enables a lower-bound estimate of the wage premium due to physical differences correlated with gender. The premium can be measured only for jobs in which both men and women were employed, and, given extensive occupational segregation, the list is rather short. The difference in wages by sex of those working on piece rates in a particular job surely understates the difference across all occupations had men and women been found in all jobs.

Data on piece-rate earnings in 1895 indicate that males earned on average 30 percent more than did females (i.e., the wage ratio was 0.77), when the piece rate was identical for both, and when both worked at the same job, in the same factory. The average ratio of female to male earnings for time-rate work in the factories sampled was about 0.60 in the 1895 report, thus the difference in physical product accounts for 23 percentage points (1.0 − 0.77) out of the original 40. As desk jobs have replaced manual labor, the returns to gender-specific differences such as strength must have decreased, and the piece-rate data give one measure. A variable for the decrease in strength with advances in technology and the replacement of white-collar for blue-collar labor, could well add another 0.10 to the 0.215 figure in Table 1, Part C, bringing the total change to 0.315.

The difference in the log of the ratio of female to male earnings in 1970 and 1890 is 0.264 using the data in Table 1; it increases to 0.360 when the 1970 figure is corrected for hours of work among full-time workers and to 0.392 when the actual data (not those constructed for 1970) are corrected for hours. The three factors in Part C account for a substantial share of the change—from 55 to 81 percent, and the addition of a factor to measure the declining return to strength increases the percentage. Of the 0.264 figure, experience accounts for 36 percent, education for 54 percent, and home time for −9 percent.

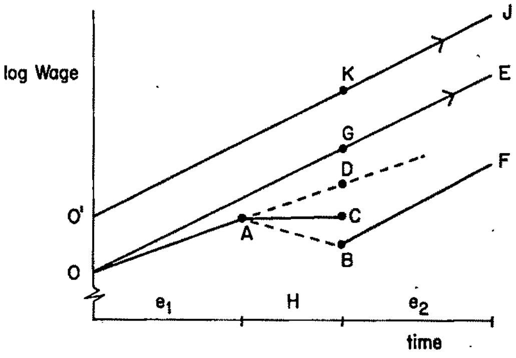


FIGURE 1. THE EFFECT OF LABOR FORCE INTERMITTENCY ON EARNINGS

The historical interlude highlights several problems in measuring discrimination and charting its course. Changes in the residual over time may suggest that discrimination increased, a conclusion that is hard to accept. Women in the 1890 labor force had little experience and education, and they were constrained in their occupational choice. Over time, their characteristics and choices expanded becoming more similar to those of men, and it is these changes that should signal a lessening, not a strengthening, of discrimination.

III. Recent Perspectives on Measuring Discrimination: An Alternative Approach

Most earnings decomposition analyses depict earnings functions as in Figure 1. An age-earnings profile for the typical full-lifetime labor force participant is given by $O'J$. It reflects earnings capacity at each level of experience and, thus, rises continuously with age. Intermittent workers generally have a different profile. The slope with respect to initial experience (e_1) is smaller, rising to level A. Earnings are definitionally zero during the period of intermittency (H), and the reentry wage (B) is lower in real terms than the wage just prior to leaving the labor market (A). Thus the total loss in wages caused by intermittency is segment (BK), the difference between the reentry wage (B) and the wage of a continuous worker. The gap can be divided into four segments: BC is the direct depreciation of skills due to their atrophy; CD is lost wages caused by lost seniority; DG is additional earnings due to extra on-the-job training obtained by those with full-lifetime work expectations; and GK measures additional earnings attributable to extra or more marketable schooling for those planning to specialize more in career than home activities.

Discrimination studies typically estimate the female earnings equation $OABF$ and compare it with the male earnings equation $O'J$ to derive the discrimination measure specified above. The distance BD is taken to measure the difference in earnings between the intermittent and the continuous worker; (BD/BK) is then the explained portion of the wage gap while (DK/BK) is the residual portion. But according to the argument just presented, the unexplained portion, and thus the measure of discrimination, may be upwardly biased, because the expectation of intermittency affects the earnings profile but is not included.

The bias was alluded to in Polachek (1975a) and Gary Becker (1985) as resulting from differences in work effort caused by the division of labor and intermittent labor force participation. An explicit consideration is given in Polachek (1975b), on which we base the empirical work below. In that study, the human capital measure incorporates expected lifetime work and is then embedded in a wage regression on pooled male-female data. The change in magnitude of a dummy sex-variable coefficient, after account is taken of human capital, measures the significance of human capital in explaining wage differentials by sex.

The approach is based on Yoram Ben-Porath's (1967) life cycle model of human capital accumulation. Ben-Porath analyzed human capital accumulation by equating marginal cost and marginal gains annually for particular earnings functions. Marginal gains are the present value of the income stream generated from an additional unit of human capital. Marginal costs, while not directly observable, can be derived from the earnings function. Specifically, annual investment net of depreciation (I_n), can be obtained as the time rate of change of earnings (y) divided by investment returns (r): $I_n = r^{-1}(dy/dt)$. For the typical quadratic function, $\log_e y_t = y_0 + rS + \beta_1 t + \beta_2 t^2$, where S = years of schooling, net investment can be measured as $I_n = r^{-1}(\beta_1 + 2\beta_2 t)$ in "time-equivalent" terms (see Jacob Mincer, 1974). In "dollar" terms, $I_n^d = I_n \exp(y_0 + rS + \beta_1 t + \beta_2 t^2)$. Adding annual depreciation of skills to net investment produces annual "gross" investment. Equating gross investment with marginal gains yields the solution for the marginal cost of investment.

We posit that men and women have the same marginal cost function, under the assumption of comparable innate skills and ability. We derive female marginal gains, equate them to the male (and by assumption the female) marginal cost to obtain female gross investment, and then apply male depreciation rates to get net investment at each age level. Summing aggregate net investment over the life cycle yields the "expected female human capital stock." Differences in male and female expected human capital stock are therefore generated only from gender differences in lifetime labor force participation.

To measure the importance of accumulated human capital as a determinant of market wage differentials, we regress earnings on the expected human capital stock for the individual: $y_i = \alpha_0 + \alpha_1 E(CAPSTOCK) + \alpha_2 SEX + \Sigma\beta_i V_i + \varepsilon_i$, where y_i is earnings of individual i, $E(CAPSTOCK)$ is the expected capital stock computed as above, and the V_i are other individual characteristics. The coefficient α_1 can be interpreted as a rate of return to the expected capital stock. The variables that constitute V, such as number of children, occupation, and industry, affect postschool investment but are not directly used in its computation. Their coefficients, when $E(CAPSTOCK)$ is included, can be interpreted as the effects of deviations from expected labor market activity. The SEX variable is a dummy (1 = female), and its coefficient is the dollar male-female earnings differential adjusted for each of the independent variables.

TABLE 2—EARNINGS FUNCTIONS FOR MALE AND FEMALE WORKERS, 1980 (Dependent Variable = Annual Earnings)

A: Pooled Samples of Males and Females, Married and Single

	Married		Single	
	(1)	(2)	(1)	(2)
Constant	−19656.7	−7994.0	−18117.2	−7389.8
	(72.03)	(38.20)	(30.88)	(38.20)
EDUC	1207.4		866.17	
	(94.78)		(29.34)	
EXP	472.60		478.62	
	(41.53)		(18.68)	
EXP^2	−6.94		−8.04	
	(28.99)		(13.09)	
HOURS	166.42	157.88	125.44	131.71
	(47.85)	(45.33)	(15.28)	(15.54)
WEEKS	212.78	227.86	247.84	251.69
	(68.68)	(73.73)	(35.34)	(34.73)
E(CAPSTOCK)		0.115		0.086
		(101.3)		(25.96)
SEX	−7741.3	−1717.2	−2424.1	−1377.9
	(96.31)	(16.84)	(14.64)	(7.91)
R^2	.43	.43	.13	.30
NOBS	62,129	62,129	7,142	7,142

B: Males and Females Separately

	Males		Females	
	(1)	(2)	(1)	(2)
MRST	2030.4	389.19	−1516.44	−578.29
	(11.27)	(2.16)	(15.62)	(6.04)
E(CAPSTOCK)		0.110		0.080
		(81.90)		(54.00)
R^2	.28	.25	.41	.40
NOBS	41,164	41,164	28,107	28,107

Sources: 1/1000 Public Use Sample of the 1980 U.S. Population Census.

Notes: Absolute values of *t*-statistics are in parentheses. $SEX = 1$ if female; *E(CAPSTOCK)* is defined in the text; other variable definitions are as in Polachek (1975b); Part A, married; The sample consists of married males and females. Control variables are region, city size, occupation, industry, and years married. Part A, single: The sample consists of white, never married males and females. The *E(CAPSTOCK)* variable accounts for the probability of becoming married. Part B: $MRST = 1$, if married; col. 1 also adjusted for *EDUC*, *EXP*, EXP^2, *HOURS*, and *WEEKS*; col. 2 also adjusted for *HOURS* and *WEEKS*.

The estimation results are given in Table 2; Part A includes separate estimations for married and single individuals. The most relevant coefficients are the values obtained for *SEX*. Two columns are presented, one in which adjustments are made for traditional human capital variables (age, education, weeks, and hours of work), and one which incorporates the expected capital stock variable [*E(CAPSTOCK)*]. In Polachek (1975b) the coefficient on *SEX* was −3032, the dollar difference in earnings using the *1960, 1/1000 U.S. Population Census Sample*. But when appropriate account was taken of life cycle differences in labor force expectations, the male-female earnings differential was reduced to −325. Fully 89 percent = [(3032 − 325)/3032] of the earnings differential could be explained when accumulated human capital, defined above, was incorporated. Note that this underestimates the true figure because the 3032 differential already adjusts for age, education, years married, and hours of work. For 1980 the figure is 78 percent = [(7741 − 1717)/7741]. When one adjusts by occupation, industry, location, length of marriage, and number of children the differences are further diminished. For single males and females, two groups with relatively small earnings differences, 38 percent (for 1960) and 43 percent (for 1980) of the remaining differentials are explained.

To lend credence to these results, a similar computation is performed for marital status differences within each of the male and female samples. In the 1960 estimation, 82 percent of the $3000 earnings premium received by married males was explained by life cycle differences in labor force participation between married and single individuals. A somewhat smaller percentage, 75 percent to be exact, of the $625 premium earned by single women was explained by their lifetime labor force participation. For 1980, these two figures are 81 and 62 percent (see Table 2, Part B). The smaller explanatory power of *E(CAPSTOCK)* in the female data probably owes to the problem of controlling for intensity of work among individuals with greater home responsibilities.

IV. Summary Remarks

A concept, termed the "residual," has become the accepted standard to measure the degree to which economic outcomes in the labor market result from discrimination. Our two perspectives suggest several flaws in the concept. The historical perspective demonstrated that as the ratio of female to male wages grew over the last century the residual remained constant. But the increase in the residual as a percentage of the wage gap does not necessarily signal increased discrimination; instead, it indicates that it is easier to explain differences in wages by sex a century ago than today. The reasons are clear. The female labor force around 1900 was a relatively homogeneous group—young, unmarried, ill-educated, and inexperienced. A complex set of forces, including an increased value of acquirable skills and a change in structure of labor demand, led to an expanded and more heterogeneous female labor force. The expansion in the acquired characteristics of female labor force participants was due, at least in part, to reduced constraints on women, but the residual appraises discrimination by differential rewards, not differential characteristics.

It was the heterogeneity of the female labor force, in cross section and over its life cycle, that motivated the second perspective, a direct measure of expected lifetime human capital. To obtain the measure, an implicit model of expectations was invoked that assumes individuals form expectations on the basis of their elders' experiences. That model of expectations was probably more accurate for 1960 than 1980 (see Steven Sandell and David Shapiro, 1980, for direct evidence), and the possible underestimation of the capital stock could account for the smaller explanatory power of the expanded human capital model in 1980. The empirical results for both years reveal a rather large percentage of the difference in earnings between men and women results from different expectations of future employment and thus from different expected stocks of human capital. But we caution that it would be incorrect to conclude that the findings indicate a virtual absence of discrimination.

REFERENCES

Becker, Gary, "Human Capital, Effort, and the Sexual Division of Labor," *Journal of Labor Economics*, January 1985, *3*, S33–S58.

Ben-Porath, Yoram, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, August 1967, *75*, 352–65.

Blau, Francine and Beller, Andrea, "Trends in Earnings Differentials by Sex: 1971–1981," paper presented at the American Statistical Association Meetings, Chicago, 1986.

______ **and Ferber, Marianne A.,** "Discrimination: Empirical Evidence from the United States," *American Economic Review Proceedings*, May 1987, *77*, 246–250.

Corcoran, Mary and Duncan, Greg. J., "Work History, Labor Force Attachment, and Earnings Differences Between the Races and Sexes," *Journal of Human Resources*, Winter 1979, *14*, 3–20.

Daymont, T. N. and Andrisani, P. J., "Job Preferences, College Major and the Gender Gap in Earnings," *Journal of Human Resources*, Summer 1984, *19*, 408–28.

Eichengreen, Barry, "Experience and the Male-Female Earnings Gap in the 1890s," *Journal of Economic History*, September 1984, *44*, 822–34.

Filer, Randall K., "Sexual Differences in Earnings: The Role of Individual Personalities and Tastes," *Journal of Human Resources*, Winter 1983, *18*, 82–99.

Goldin, Claudia, "The Historical Evolution of Earnings Functions and Occupations," *Explorations in Economic History*, January 1984, *21*, 1–27.

______, "The Earnings Gap Between Male and Female Workers: An Historical Perspective," NBER Working Paper No. 1888, 1986.

______, "Assessing Various Theories of Discrimination by Sex Using Manufacturing Data," in progress.

Leonard, Jonathan S., "Antidiscrimination or Reverse Discrimination: The Impact of Changing Demographics, Title VII, and Affirmative Action on Productivity," *Journal of Human Resources*, Spring 1984, *19*, 145–74.

Madden, Janice, "The Persistence of Pay Differentials: The Economics of Sex Discrimination," *Women and Work*, 1985, Vol. 1, 76–114.

Mincer, Jacob, *Schooling, Experience, and Earnings*, NBER, New York: Columbia University Press, 1974.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, October, 1973, *14*, 693–709.

Polachek, Solomon, (1975a) "Potential Biases in Measuring Male-Female Discrimination," *Journal of Human Resources*, Spring 1975, *10*, 205–29.

______, (1975b) "Differences in Expected

Post-School Investment as a Determinant of Market Wage Differentials," *International Economic Review*, June 1975, *16*, 451–70.

Sandell, Steven and Shapiro, David, "Work Expectations, Human Capital Accumulation, and the Wages of Young Women," *Journal of Human Resources*, Summer 1980, *15*, 335–53.

Voos, Paula, "Estimates of Wage Discrimination Against Women Based on the Direct Measurement of Male and Female Productivity," unpublished paper, University of Wisconsin, January 1986.

Race and Poverty: A Forty-Year Record

By JAMES P. SMITH AND FINIS WELCH*

Thirty years ago, Gary Becker in his now classic work, "The Economics of Discrimination," sparked renewed interest in an economic analysis of racial income disparities. The volumes of research papers that built on Becker's contribution over the last three decades added a great deal to what we know about the reasons for the wide income differences between the races. One reason was the emergence of several large scale micro data sets of which the 1960 census was the first. Today, analysis is based not only on the 1980 census file but also on several longitudinal data sets best represented by the *Panel Study of Income Dynamics* and the "Parnes" *National Longitudinal Surveys*. Ironically, it is the release of micro data files from two pre-Becker data sets that appears to offer the greatest potential for answering the important questions that remain.

In this paper, we use these two data sets—the 1940 and 1950 census files—in combination with the three subsequent census files to describe long-run trends in black poverty. We begin by describing purely labor market developments, but supplement that depiction with a broader look at events that impacted on the black family. The paper concludes with an examination of the downside of black economic progress—the increasing disengagement of many black men from the labor market.

I. Good Jobs and Bad Jobs

What has happened to the jobs black workers have been able to obtain over the last forty years? The 1940–80 census files point to a very impressive rise in the relative economic status of black workers over the last forty years. For example, in 1940, the average wage of a black man was only 43 percent of the average of white men. By 1980, it was 73 percent. Since 1940, black male wages have increased 52 percent faster than those of white men.[1] Since the labor market gains achieved by black women far exceed those of black men, bringing women in only adds to the impression of progress. In 1940, working black women earned 40 percent of the weekly wage of working white women. Forty years later, racial wage parity among women had been achieved with black women earning 99 percent of the wage of white women.[2]

So far we have talked about progress as measured for the typical black worker. There is a legitimate concern that these gains were heavily skewed with some black workers receiving the bulk of the benefits, leaving large numbers of black workers behind. This concern is heightened by a growing fear that other forces have increasingly blunted the translation of improving labor market status into higher levels of well-being for the black family. The issue of the distribution of black labor market progress is addressed in Table 1. Building on the simplicity of the poverty line, we divide all workers into three weekly wage classes—the poor, the middle class, and the affluent.[3]

Using this simple three-way division, white men had a lock on the good jobs in the 1940 labor market. Coming out of the depression, however, even 29 percent of working white men had jobs that placed them in poverty.

*The Rand Corporation, Santa Monica, CA 90406–2138, and Unicon Research Corporation, respectively. We acknowledge the expert assistance of Judith Fischer and Eanswythe Leicester. Grants from NICHD to the RAND Corporation and from the Olin and Sloan Foundations to the RAND Corporation and Unicon Research Corporation are gratefully acknowledged.

[1] See our 1986 report for a detailed analysis of trends in black-white male weekly wages.

[2] The racial wage ratios for women across the 5 decades are 40.3, 67.7, 61.2, 82.7, and 99.3.

[3] To do so, we had to define two income thresholds in each census year, a poverty threshold below which are the poor and an affluence threshold above which are the affluent. For details on how this was done, see Smith (1986).

TABLE 1—INCOME GROUP STATUS OF WORKERS[a]

	1980	1970	1960	1950	1940
White Men					
Poor	11	9	12	17	29
Middle Class	61	68	64	61	42
Affluent	28	23	24	22	29
Black Men					
Poor	21	25	38	49	76
Middle Class	67	70	59	48	21
Affluent	11	5	3	3	3
White Women					
Poor	41	39	44	40	57
Middle Class	54	56	52	54	34
Affluent	5	5	4	6	9
Black Women					
Poor	41	51	73	74	93
Middle Class	54	46	27	23	5
Affluent	5	3	2	3	2

[a] Numbers are the percent of workers of each race-sex group in each income stratum.

But the situation for our other three demographic groups was far worse. Three-quarters of employed black men worked in jobs that confined them within the ranks of the poor. Only one in five black men earned middle-class wages. Similarly, if they had to rely on their wages alone, 57 percent of working white women had jobs with wages below our poverty cutoff. Working black women were at the bottom with an astonishing 93 percent earning poverty level wages. On the other side of the track, the elite jobs in 1940 resembled an exclusive white male club. While 29 percent of white men were members, our other three demographic groups lagged quite far behind.

The subsequent changes have been dramatic. In 1980, fully 21 percent of working black men still languished within the poor underclass, a reminder (if any of us needed it) that many black men remained left out and left behind. However, the real story for black men over the last forty years has been the spectacular reductions in the ranks of the black working male poor alongside the emergence of a black male middle class. The timing of these changes is also of interest. Ironically, many of the victories in the war on poverty apparently took place five years before the war was officially declared. The relative performance across these decades described in Table 1, with 70 percent of the poverty reduction taking place before 1960, offers important clues about the real contributors to long-run predictions in black poverty.

Nowhere are these changes more dramatic than when we focus on those with jobs within the contemporary economic elite. For the first time in American history, a sizable number of black men are working in jobs that are better than those of middle-class male whites. During the last twenty years, the odds of a black working man having an affluent job have almost quadrupled.

These labor market gains were even larger among working black women. In the course of little more than a generation, the labor market faced by black women has been transformed beyond recognition. Black working women have persistently moved out of the poverty level jobs into those corresponding to the white male middle class. In contrast to the situation among black men, where labor market progress slowed during the last decade, the strides made by black women continued unabated. By 1980, almost 6 in 10 working black women had a job with a wage that met the minimum requirement for the working white middle class. Once again, balancing that progress are the figures showing the still quite large contemporary black female working poor. In 1980, 4 in 10 black women still had jobs that labelled them as the working poor.

What were the principal long-run causes of this substantial improvement in the labor market position of black working men and women? If our research has developed a single theme on this question, it is that long-run movements in racial income disparities largely mirror secular trends in racial skill differences. The two central skill dimensions were schooling—both in additional years of schooling and the quality of schools—and the historic migration of blacks from the rural South to the northern cities. These two factors combined were the dominant forces accounting for the changes documented in Table 1.[4]

[4] See our earlier study and Smith (1986) for details.

Finally, the impact of affirmative action has been mixed. Affirmative action had no significant long-run impact, either positive or negative, on the male racial wage gap. Among males, affirmative action did have a pro-skill bias, permanently raising the incomes of young black college grads. As a result, affirmative action, while irrelevant to reductions in black poverty, did contribute to the recent growth of the black male elite during the 1970's. In contrast, black women, meeting the requirements for two quotas, were the primary beneficiary of affirmative action. Affirmative action played an important role in raising wages of black women and partly accounts for the continual improvement in the economic status of black women during the 1970's.

II. Black Family Poverty

Trends in labor market wages of black men and women represent only a first step in understanding the changes in black poverty that have taken place. There remains a real question of whether these labor market gains touched the average black family. To determine long-run trends in numbers of poor families since 1940, Table 2 separates families into three income classes for all census years between 1940 and 1980.[5]

Turning to our main focus, the race dimension, this table simultaneously illustrates the persistence of black family poverty, the growth of a black middle class, and, more recently, the emergence of a black elite. Poverty was pervasive in the 1940 black community. In 1940, 70 percent of black families were destitute, with little hope that their lot or that of their children would soon improve. The black middle class of 1940 was correspondingly small, counting among its members only one in every four black families.

The important development over the last forty years has been the emergence of a black middle class. From a very dismal starting point, progress in reducing black family poverty has been impressive. By 1960, for the first time in American history, the median black family was not poor. However, fully 30 percent of black families in 1980 were still part of the black underclass. But placed in historical perspective, such figures still represent enormous progress toward eradicating black poverty. There was, however, a slackening of this progress in reducing black family poverty during the 1970's, while at the same time the black elite continued to expand. Since 1960, there has been an increasing disparity between the continued improvement in black labor market outcomes and a stagnation in the economic welfare of the black family.

TABLE 2—INCOME GROUP STATUS OF FAMILIES[a]

	1980	1970	1960	1950	1940
All Families					
Poor	11	11	15	22	34
Middle Class	63	66	62	50	40
Affluent	26	23	23	28	26
White					
Poor	9	9	12	18	31
Middle Class	61	66	63	52	42
Affluent	30	25	25	30	27
Black					
Poor	30	32	48	54	71
Middle Class	59	59	49	42	26
Affluent	11	9	6	4	3

[a] Numbers are the percent of the population in each income stratum.

Not all families participated equally in this economic progress. Particular concern has been expressed about the plight of the growing numbers of black families headed by women. To address this issue, Table 3 presents our three-way separation of income class for intact (both spouses present) and female-headed families.

The serious economic problems faced by many black families can hide an often overlooked and numerically larger group of husband-wife black families who are doing quite well. In 1940, even their poverty rates hovered about 70 percent, attesting to the pervasiveness of black poverty at that time. But among all demographic groups, the subsequent gains were largest among intact black families. Their poverty rates were cut in half between 1940 and 1960, and economic pro-

[5] This section is based on Smith (1986).

TABLE 3—INCOME GROUP STATUS OF FAMILIES
(In percent)

	1980	1970	1960	1950	1940
Intact Families					
White					
Poor	6	7	10	17	30
Middle Class	64	67	64	52	41
Affluent	30	26	26	31	29
Black					
Poor	15	21	39	49	69
Middle Class	68	69	54	44	27
Affluent	17	9	7	7	4
Female-Headed Families					
White					
Poor	30	32	34	42	41
Middle Class	62	58	52	40	39
Affluent	8	10	14	18	20
Black					
Poor	53	58	69	76	81
Middle Class	44	40	29	21	17
Affluent	3	2	2	3	2

gress continued unabated up to 1980. Only 15 percent of intact black families had incomes below the 1980 poverty cutoff.

The last twenty years are especially remarkable for the spectacular rise in the fraction of husband-wife black families who joined the affluent class. In such families, affluence rates doubled with much of the rise taking place during the 1970's. Today almost one in five black intact families are members of the economic elite.

Unfortunately, there is a full complement of bad news to tell as well and much of it concerns female-headed families. In 1940, 18 percent of black families had a female head compared to a 10 percent rate among whites. The subsequent changes were relatively small until 1960, a year in which one in five black families were headed by women. By 1980, four in every ten black families were female headed compared to a 12 percent rate among whites.

Families headed by women were always more likely to be poor. However, the salient secular trend is an ever-expanding divergence between the economic status of female-headed families from that of intact families. Average incomes of female-headed families have simply failed to keep up with economywide income growth so that their situation has worsened relative to other families. Even today, almost a third of white female-headed families are poor and more than half of black families are. This divergence is best illustrated by the sharply declining fraction of female-headed families with sufficient incomes to gain membership among our affluent class.

The increasing concentration of the black poor in female-headed families has legitimately become a source of major public policy concern. Close to 60 percent of all blacks below the median black family income were members of families with women as heads. At the other extreme, only 6 percent of the richest 5 percent of black families were female headed. The rising fraction of black female-headed families is the primary cause of a number of contemporary dimensions of black poverty. It largely accounts for the feminization of black poverty (with 69 percent of poor black adults being women). Similarly, 70 percent of poor black children now live in female-headed families compared to a 30 percent rate twenty years ago. Knowing whether a black family is female headed has become an increasingly accurate predictor of black economic status.

What role then has the breakup of the black family played in stemming the long-term reduction in black poverty? As for many race issues in this country, opinions have polarized into extreme positions. At one extreme are those who argue that except for the continuing breakup of black families, the problem of black poverty would be over. Others dismiss the black family as a cause, partly because it raises questions about the efficacy about some existing social programs.

How much of existing black poverty can legitimately be assigned to the problems of the black family? In a recent paper, Smith (1986) answered this question by calculating what the poverty rate among black families would have been if the incidence of black female-headed families in 1980 was the same as it was in the white population.[6] The result

[6] The details of the methodology are provided in Smith's earlier paper.

of that calculation shows that the quantitative truth lies between the polar opinions. Instead of the actual rate of 29 percent, 23 percent of black families would be poor if we could magically impose on the black population, the white rates of female headship. To put it one way, one-third of existing racial differences in poverty are due to the instability in the black family, a not-insignificant number. However, to put it another way, even if the black family issue were entirely resolved, black family poverty rates would still far exceed those of whites, and almost a quarter of black families would remain mired in poverty. Clearly other factors are also playing a central role, an issue we return to below.

III. The Downside of Black Progress

To this point, we have written with an optimistic pen by highlighting a forty-year record of black economic progress. This optimism must be tempered by a recognition of the high levels of black poverty that persist, a recent slowdown in the reductions in black poverty, and some real concerns that large segments of the black community are no longer participating in this progress because their involvement in the labor market has ended.

What are the reasons for the slowdown in decreasing black poverty? One frequently mentioned candidate, a cessation of the long-run trends of improving black labor market skills and wages, turns out not to be plausible. It is important to remember that labor market progress in narrowing the wage gap continued during the 1970's. For example, black working men earned 73 percent as much as whites in 1980 compared to 67 percent in 1970. But three events have started to blunt the translation of the still-improving black labor market skills into a higher standard of living for black America: the accelerating breakup of the black family, rising rates of black unemployment, and a slowdown in American economic growth.

In spite of the long-term improvement in their labor market opportunities emphasized here, an increasing number of middle-aged black men in recent years have dropped out of the labor force and many young black men faced increasing difficulties in finding their first full-time job. Indeed, a common reaction to our work is that the deterioration in the employment side of the black labor market is so severe that it renders false the positive implications derived from the wage side. Real and growing problems exist in the employment dimension, but they are often mischaracterized, assigned to the wrong group of black workers, and thus far lack a convincing empirical explanation.

TABLE 4—ACTIVITY STATUS OF BLACK MEN

	1980	1970	1960	1950	1940
18-Years Old					
SEM	79.3	79.5	78.6	82.1	78.3
UOJ	20.6	20.4	21.4	17.9	21.7
24-Years Old					
SEM	71.8	78.9	80.2	86.2	82.7
UOJ	28.2	21.1	19.8	13.8	17.3
35–36-Years Old					
SEM	79.7	86.3	82.9	86.5	85.3
UOJ	20.3	13.7	17.1	13.5	14.7

Note: SEM = School, Employed, Military; and *UOJ* = Unemployed, Out of the Labor Force, Jail.

To see this, Table 4 lists the activity status of young black men. Because conventional labor force statistics give a misleading portrait of activities of young men, we have divided activity status into two groups. *SEM* (the "good" activity) includes schooling, work, and the military, while *UOJ* (unemployment, out of the labor force, and jail) are the bad ones.

It may come as a surprise to many that there exists no negative secular trend for black male teenagers. Throughout the last forty years, roughly one in five black male teenagers was confined to unproductive activities. In fact, the principal secular trend among black male teenagers was within the *SEM* group, where the fraction in school increased by 35 percent largely at the expense of the fraction at work. Since this is certainly a positive development, there is nothing on the employment side of the labor market that counteracts the positive wage story for black teenagers.

Although the seeds of the problem are certainly sown earlier, the secular deterioration, instead, is concentrated among somewhat older black men. More important, it is confined entirely to the 1970's: 28 percent of 24-year-old black men are in the unproductive group in 1980, a jump of 7 percentage points since 1970. In this age group in 1980, one in nine black men were unemployed, another one in nine were out of the labor force, and one in twenty were in jail. Similarly, one in five black men aged 35–36 are now assigned to our unproductive activities class, with half of them completely absent from the labor market.

The prevalence rates in Table 4 tell only part of the story. The real question involves dynamics; the extent to which black men in the unproductive activities at any point are able to make their way out. Table 5 represents a crude attempt at capturing some of these dynamics by listing transition rates between activities observed one year apart.[7]

There are three principal conclusions we draw from Table 5. First, there are a large group of black male workers with reasonably stable employment, as indicated by their eventual 94 percent transition from employment to employment. Second, there exists a smaller group of black workers (running from 20 to 30 percent) with far less stable employment. Within this group, the currently unemployed and those out of the labor force behave quite differently. Among the unemployed, about half have jobs one year later, a rate that varies little with age. The one-third still found unemployed attests to the severity of the problem. Third, it is the out of the labor force transitions that isolate the new employment problem. By the time these black men are in their mid-30's, at least as measured across a yearly time frame, many seem to be lost to the labor market. At that age, 70 percent remain out of the labor force a year later.

TABLE 5—LABOR FORCE TRANSITION OF BLACK MEN

	Ages		
	16–18	22–27	32–38
Employed			
Unemployed	13.7	10.4	3.3
Out of Labor Force	4.9	3.0	2.3
SEM	81.4	86.6	94.4
Unemployed			
Unemployed	34.6	39.7	39.2
Out of Labor Force	6.3	10.6	7.8
SEM	59.0	49.7	53.0
Out of Labor Force			
Unemployed	16.7	19.5	7.5
Out of Labor Force	33.3	53.7	71.2
SEM	50.0	26.9	21.3

Note: SEM = School, Employed, Military.

In our view, the major unsettled research question on race centers around the reasons for the growing fraction of persistently disengaged black men from the labor market. The competing hypotheses are not new, but a convincing case has not been made on either side. William Wilson (1986) and others have emphasized the demand side, claiming that the substantial restructuring of the economy in the last fifteen years eliminated jobs that were disproportionately held by inner-city blacks. The supply-side counterpoint, swayed in part by the coincidence in timing, directed attention towards the host of social programs now subsumed under the popular label, the "safety net." It is argued that these programs are an attractive alternative to work for many black men whose market rewards are meagre. The political rhetoric surrounding this issue is obviously intense, but it has not been matched by scientific precision in settling the question.

There is one dimension of the employment on which more can be reasonably said at this time—the impact of the slowdown in American economic growth. The sustained and rapid growth of the post-1940 American economy carried with it impressive benefits that helped blacks and whites alike. For

[7] The transition rates in Table 5 were derived from matched March *Current Population Survey* tapes from 1977 to 1984. Because of their sample sizes, the data were pooled across years and across the age intervals used in Table 5. Because of the sample design of the *CPS*, the activity "in jail" is not represented. Since the *CPS* is a "roof-top" survey that repeats observations on the same location, the transitions in Table 5 are undoubtedly affected by differential geographical mobility by transition.

example, inflation-adjusted incomes of white men expanded two and one-half-fold since 1940. Thus, the whites to whom one compares blacks in 1980 were far wealthier than the whites who represent the contrast group in 1940. According to our estimates, 45 percent of the reduction in black male poverty since 1940 was due to economic growth and the remaining 55 percent to the expanded black labor market skills. Economic growth and improving black labor skills, principally through education, go hand in hand as the key weapons history identifies as eradicating black male poverty.

Economic growth has had its dark side lately. Between 1970 and 1980, real incomes grew by less than 3 percent, one-tenth of the growth achieved during the previous decade. The virtual absence of the real income growth during the 1970's carried a terrible price in limiting reductions in the ranks of the black poor. The key problem for blacks in the 1970's was not that their situation did not improve relative to whites, it was that, for the first time, the whites they were being compared to were no better off in 1980 than in 1970.

What would black family poverty rates have been in 1980 if two disquieting events of the 1970's—the absence of economic growth and the accelerating breakup of the black family—had not occurred?

If economic growth had been maintained at the rapid rate of the twenty years between 1950 and 1970, black family poverty would have been 5 percentage points lower with 24 percent of black families counted among the poor. If we then also maintain black female headship rates at their 1970 levels, black poverty rates would decline further to 21 percent. In combination, these two factors fully account for the fact that black family poverty rates did not continue to decline at historical rates during the last decade.

IV. Conclusion

In this paper, we have described a forty-year record of progress in reducing black poverty. The optimism that should have resulted from that progress was tempered by three other events of the 1970's: the accelerating breakup of the black family, the growing numbers of black men with little contact with the labor market, and a continued slowdown in economic growth. As a result, we have entered a new racial era in America in which black labor market progress no longer guarantees improvements in economic welfare, especially for the black underclass.

REFERENCES

Becker, Gary, *The Economics of Discrimination*, Chicago: University of Chicago Press, 1971.

Smith, James P., "Poverty and the Family," paper presented for the Institute for Research on Poverty conference on Minorities and Poverty, December 1986.

Smith, James P., and Welch, Finis, *Closing the Gap*, R–3330–DOL, Santa Monica: Rand Corporation, 1986.

Wilson, William J., "Social Policy and Minority Groups," paper presented at the Institute for Research on Poverty conference on Minorities and Poverty, December 1986.

Debt, Capital Flows, and LDC Growth

By ANNE O. KRUEGER*

Since 1982, newspaper headlines have called attention to the "debt crisis," or debt problem of the developing countries (LDCs). Initially, most analysts believed that debt-servicing problems would be temporary and that creditworthiness and more normal growth of most countries would be restored in a period of at most several years. However, events have demonstrated that this initial assessment was optimistic. This paper focuses on the debt problem and the reasons for its persistence. That in turn permits an analysis of the policy challenges that must be met for a resumption of LDC growth and a return to more normal capital flows.

Until the 1980's, virtually all analysts viewed capital flows the LDCs and the apparent shift from official to private flows as a healthy development. Capital flowing from rich countries with relatively low rates of return on investment and high savings rates to poor ones with higher rates of return on investment and lower savings rates was seen as an efficient allocation of world resources. These flows were thought to benefit all: certainly labor in poor countries would benefit and per capita incomes would rise more rapidly with capital inflows to poor countries than if investment were constrained by domestic savings rates.

Moreover, it was thought that with high real rates of return, capital flows would be self-financing: returns on investment would cover servicing obligations. An important question for policy is whether this basic view remains correct. While a full discussion of capital flows would consider their breakdown between debt and equity financing, that issue is secondary to analysis of the current situation and is ignored here.

I. Evolution of Capital Flows to LDCs

After World War II, the conscious push for development began when there were virtually no long-term private capital flows. Although the private international capital market revived in the late 1950's, for LDCs capital flows remained almost entirely official. This remained the rule in the 1960's with the notable exceptions of a few dramatically successful countries. They sharply increased their rates of growth, and returns on investment, by moving to integrate their economies with the world economy, borrowing to supplement domestic savings and, in some cases, encouraging larger flows of direct foreign investment.

Observers of the progress of the LDCs regarded this development as a hopeful sign: it appeared that at low levels of income, there was an initial period during which infrastructure investments in transport, communication, education, and the like could not be financed on the private market but that, with continuing growth, development could readily proceed with reliance on private international capital markets.

After the oil price increase of 1973, most oil-importing LDCs had large current account deficits. Simultaneously, oil exporters had large current account surpluses, which were used at first to accumulate short-term liquid claims in large commercial banks in developed countries. Given the banks' favorable experience with their lending to the rapidly growing outer-oriented LDCs, it is not surprising that many LDCs could finance their current account deficits through the private banks.

[†] *Discussants*: Manuel Guitian, International Monetary Fund; Armeane Choksi, World Bank.

*World Bank and Duke University, Durham, NC 27706.

While most LDCs borrowed, the underlying economic rationale differed widely among countries. In some instances, LDC current account deficits financed newly profitable high-return investments which were undertaken because the countries had increased incentives for production of tradables. Those countries quickly reduced current account deficits and resumed "normal" borrowing. Other countries borrowed primarily to maintain their preexisting patterns of consumption, investment, and government expenditures. Some of them even maintained pre-1973 energy prices while incurring large current account deficits. Some countries' failure to adjust was so extreme that they met debt-servicing problems in the latter half of the 1970's. There were debt crises and debt reschedulings, although normally not for more than a few in any year.

For many oil-importing LDCs, however, adjustment to the deterioration in their terms of trade was partial, and borrowing to sustain the rate of investment while consumption had increased was frequent. The world economic environment obscured any underlying problems. It will be recalled that the 1976–79 period was characterized by largely unanticipated worldwide inflation; until at least 1978, most lending was at fixed interest rates, and the resultant *ex post* real rate of interest was negative. Developing countries' nominal export earnings grew at average annual rates in excess of 15 percent, while the real value of outstanding debt was diminished by inflation. The result was that, despite significant new borrowing, the real value of debt outstanding in 1977 was below the level of 1972. As a percent of GNP, LDCs' debt did not start rising significantly until after 1977.

Thus, although oil-importing LDCs were borrowing large sums in the 1970's, many were borrowing to permit continuation of policies that could not have been sustained had real interest rates been realistic. The external environment in effect validated their policies. While some investments probably had relatively low rates of return, that was not a concern in an environment of negative real interest rates, except for countries where excesses were too great. Although it can be argued that a shift in worldwide conditions might have been anticipated, there was little in the aggregate statistics to suggest that the capital flows of 1973–78 were unsustainable, given worldwide inflation.

II. The Debt Problems of the Early 1980's

The shift in worldwide conditions in 1979–80 was as dramatic as that of 1973–74. Although the proportionate increase in the oil price was much smaller than in 1973, oil and energy were a much larger share of imports and expenditures. The initial response to the second oil price increase seemed to mirror the first. In the first year, worldwide inflation accelerated, oil-importing countries' current accounts swung sharply negative (or more negative), and recession set in.

But the similarity ended in 1980. First, whereas most countries at the time of the first oil price increase had sustainable current account positions and moderate levels of debt, this was not true at the time of the second. Moreover, economic activity in the industrial countries did not revive. Instead of adopting traditional Keynesian policies, the developed countries adopted anti-inflationary policy stances which resulted in a prolonged recession, reduced growth (shrinkage, by 1982) of world trade, declining commodity prices, and sharply higher nominal and real interest rates. The result was that developing countries' current account deficits rose from the $30–35 billion range of 1978–79 to $59 billion in 1980 and over $100 billion in 1981 and 1982. Oil-importing LDCs' current account deficits rose from 2.2 percent of GNP in 1978 to over 5 percent in 1981 and 1982.

The consequent increase in debt was staggering. Although many countries financed part of their deficits by reducing reserves and borrowing short term, long-term debt rose from $359 billion in 1979 to $552 billion in 1982. The increase in real debt was even more pronounced as export unit values were virtually stationary from 1980 to 1981 and fell by 6 percent in 1982. Thus, the ratio of debt to exports, which had stood at 1.51 in 1978 and 1.31 in 1980, rose to 1.88 by 1983.

Simultaneously, the real interest rate relative to export prices had risen from a negative 3–6 percent in the late 1970's to a positive 16–20 percent by 1982! With inflation, too, the fraction of debt subject to variable rates rose sharply. The result was an increase in interest payments from 4.6 percent of exports in 1978 to 8.1 percent in 1982 and 8.3 percent in 1983. Stated in another way: of the 1982 current account deficit of developing countries of $85.7 billion, $49.5 billion were in interest payments. These payments were a significant cause of new debt just at a time when high real interest rates meant that the debt level was too high and should probably be cut back.

By 1982–83, many developing countries found themselves unable voluntarily to continue servicing their debt. Everything contributed: exports were shrinking because of the worldwide recession; even for countries that shifted incentives strongly toward the production of tradables, there were lags in supply; it proved difficult to achieve deep enough cuts in imports simply because of time lags and the fact that adverse shifts in the terms of trade and shrinking world markets impelled such sharp reductions; higher interest rates on a greatly enlarged debt outstanding further intensified the problem.

Although worldwide conditions affected all countries, their policy history in the 1970's significantly affected their positions. It is ironic and instructive that Mexico, an oil exporter, was the first large and highly publicized country with a "debt crisis." In Mexico's case, highly expansionary macroeconomic policy had been financed through capital inflows that seemed warranted based on her rapidly rising oil exports. Nonetheless, that macroeconomic policy was unsustainable; there could be no reasonable doubt that Mexico's debt crisis was the result of domestic economic policies.

By contrast, Brazil had in 1981 undertaken a series of measures designed to increase incentives for the production of tradeables, reduce excess demand in the domestic economy, and hence improve the current account balance. In Brazil's case, the worldwide recession meant that policy changes which would likely have been adequate for at least a few years under normal circumstances failed. For Brazil, inability to continue normal debt-servicing came in 1983.

There were many other combinations. Argentina, for example, entered the 1980's with a strong external position, little external debt, but with macroeconomic imbalances and a highly distorted trade regime. Highly expansionary macroeconomic policies in the early 1980's exacerbated these difficulties and Argentina would probably have confronted a debt-servicing crisis sometime in the first half of the 1980's even had worldwide conditions been normal. For Chile, a policy mistake (a rapidly appreciating real exchange rate which resulted in capital inflows equal to as much as 10.9 percent of GNP in 1981) was compounded by sharply deteriorating terms of trade for her major exports.

Some other countries maintained their debt-servicing obligations throughout the worldwide recession. Some, such as Korea, were in reasonable policy balance when the oil price rapidly rose, and undertook sharp policy adjustments which quickly restored external balance. Turkey was notable for having had highly expansionary monetary and fiscal stances in the 1970's, combined with a severely distorted trade regime and a panoply of controls on private economic activity. The Turkish debt crisis actually started before the second oil price increase, and culminated in a major policy reform program inaugurated in January 1980; Turkish economic performance improved greatly during the worldwide recession.

Other countries had yet different patterns: fiscal and monetary policies had been conservative in much of South Asia so that initial debt levels were very low, while trade policies had been so restrictive that these countries had largely insulated themselves from the world economy at the cost of their own growth; in most of Subsaharan Africa, pervasive controls and regulations on private economic activity, inefficient parastatals, severely overvalued exchange rates, and highly restrictive trade regimes had already extracted a high cost in the form of negative growth rates of per capita incomes in the 1970's; when the terms of trade deteriorated in the early 1980's, output and incomes fell sharply and maintenance of debt-servicing was infeasible.

Thus, the precise combination of internal and external circumstances that led to an inability to maintain voluntary debt-servicing varied from country to country and some countries avoided any interruption whatsoever. Many, however, did not.

III. Policy Response to Debt Crises

Almost always, an inability to maintain voluntary debt-servicing is also a "balance-of-payments crisis," because if borrowers were credit worthy, they could finance current account imbalances through short-term borrowing. Indeed, sizeable run-ups of short-term debt are often the precursors of balance-of-payments crises, which occur when no more short-term or other credit is available.

For most of the LDCs, the size of the necessary macroeconomic adjustment in the early 1980's was very large: current accounts which had been in deficit by 3, 5, and even 10 percent of GNP net of interest payments had to shift to 0, 2, and 4 percent surpluses —shifts of as much of 10 percent of GNP for individual countries.

The first "emergency" necessary step involved stopping increases in debt by reducing macroeconomic imbalances; in the short run, this inevitably meant curtailing imports. This, in turn, meant either reduced income levels or heightened trade restrictions. As a longer-term measure, it was essential to increase incentives for production of exportables: most LDCs' trade regimes were already distorted in favor of import-competing production, and there was relatively little scope for efficient import substitution. To effect the economically desirable shift would have entailed both a significant change in the real exchange rate and reductions in levels of protection to imports, a necessary but often politically painful step. Note that import liberalization implied that the initial restriction of imports had later to be reversed.

As already mentioned, in 1982–83, it was thought that the debt problem would be relatively short-lived. Except for 1984, however, when falling debt-service ratios resulting from rapidly rising world exports and lower interest rates seemed to validate this forecast, developments have been less favorable than anticipated.

There have been a number of disquieting phenomena. First, despite the policy reforms undertaken and the current account shifts of some LDCs, debt-service ratios have risen. Second, with a very few conspicuous exceptions, growth rates of the heavily indebted countries have been very low. Third, a key assumption in earlier optimistic assessments had been that there could be rapid growth of LDCs' exports, and yet export growth in 1985 was relatively low and, worse yet, protectionist pressures against the exports of developing countries appeared to be increasing. Fourth, some of the LDCs that initially undertook policy reforms appeared to backslide as the political pressures arising from lower incomes and slow growth became irresistible, while simultaneously it became evident that some countries were unable to carry the needed reforms sufficiently far to alter the underlying policy problems that had led to problems in the first place. Finally, even those countries whose exports grew rapidly and whose policy reforms appeared adequate were not able to attract voluntary capital inflows: net flows in the aggregate continued falling, and much of the new lending that did take place was part of rescheduling activities, rather than a resumption of voluntary flows.

To a considerable degree, these five factors constitute a vicious circle: slow growth in the OECD is in part the result of protectionist pressures against the expansion of LDC exports which in turn are used as an argument in LDCs against the needed policy reforms; the failure of policy reforms to go far enough in some LDCs contributes both to their slow growth and to the failure of their exports to grow; that in turn results in the lenders' reluctance to expand voluntary lending which in turn makes it harder to undertake policy reforms and in some cases precludes the needed resources for investment in exportables. Breaking this vicious circle is the challenge for policy.

IV. Prospects for LDC Growth and Resumption of Voluntary Lending

There are some worldwide macroeconomic preconditions that must be met if the debt problems of the LDCs are to be resolved. Chief among them is that developed country markets remain open to the exports of those countries whose policy reforms have realigned incentives sufficiently to generate the needed supply response. It is self-evident that if the LDCs cannot in any event expand their export earnings at a rate sufficient to permit a gradual reduction of the debt-service ratio over time, there can be no satisfactory resolution to the current problems. However, it is on other aspects of the policy challenge that I wish to focus here.

Any analysis of the current situation must start with an appreciation of the necessity for sufficient policy reform, the ways in which the "debt overhang" can impede realization of the benefits, and the interrelationships between these two key aspects of the situation.

Debt-servicing obligations of a significant magnitude can raise two problems for a country. On one hand, there is the necessity to earn (or save) the foreign exchange for debt-service. On the other hand, especially when debt is public, there is a public finance problem, as the government must in one way or another raise the resources.

Earning the foreign exchange for debt-service requires the shifting of incentives toward the production of tradables, primarily exportables in most cases. When incentives are shifted, investible resources need to respond to expand capacity in new lines of activity. Shifting incentives typically entails measures that will increase private rewards to more accurately reflect social returns, and also policies to reduce the "crowding out" by the public sector.

For governments to raise the appropriate resources for debt service almost always entails exactly the opposite policy response: taxes must be raised, or expenditure lowered. While lowering some forms of expenditures, and reducing inefficiencies in the public sector, can and should be part of the longer-term realignment of incentives, in the short run, increasing taxes and/or diverting public resources away from infrastructure maintenance are almost always easier.

When debt-service obligations are high, increasing public resources to service debt will be likely to reduce incentives and resources available to the private sector sufficiently to preclude the necessary investment response. In countries where debt-service obligations have jumped by 4–6 percentage points of GNP while capital inflows have fallen 3–6 percent, in the absence of inflows domestic investment rates of only 10–12 percent of GNP will be feasible. In these circumstances, it can hardly be hoped that investment will be sufficient to increase the supply of exportables sufficiently for growth.

Therein lies the policy dilemma: for countries undertaking adequate policy reforms, additional external financing may have a very high rate of return. It is quite possible that countries having undertaken the necessary reforms may find themselves caught in a low-growth, high-debt-service trap: because capital inflows have diminished while debt-service obligations have increased, domestic savings available to finance new investment is low despite high rates of return on activities newly profitable after policy reform. Simultaneously, because of high debt-service ratios and slow growth, foreign financing for these newly profitable activities is not forthcoming. In these circumstances, exports cannot grow for lack of capacity expansion, debt-service ratios will remain high, and countries may therefore remain uncreditworthy.

Symptoms of such difficulties might be several, but high real interest rates well above international levels would be one of them. Another would be increased national savings rates (before debt servicing) combined with reduced rates of investment. Finally slow growth of capacity in tradables despite high profitability would be observed.

If this "debt overhang" trap were present, the clear policy implication would be that some degree of augmented official capital flows was temporarily warranted, and would have a very high real rate of return. However, the other clear policy implication is that these official flows would be unwarranted in the absence of sufficient policy reform: failure of a country sufficiently to realign its real exchange rate and liberalize its trade regime, to align its incentive structure in ways conducive to economic efficiency, or to control its public sector deficit sufficiently to permit the realization of growth and an improved current account position would render additional capital flows no more warranted than some earlier borrowing had been.

The risk is that the resources available for truly well-founded policy packages will be inadequate, while other countries, having moved their policies in the right direction but not far enough, will increase their liabilities in ways that will prejudice the success of reform when it does finally occur. The challenge for policymakers and researchers is to determine "how much is enough" by way of policy reform, and to distinguish between cases in which a "debt overhang" phenomenon arises from cases where policy reforms have simply been inadequate to the task. Categorizing all indebted countries, regardless of the degree of policy reform, as being in the same category is clearly a policy mistake.

It is relatively simple to ascertain whether countries are undertaking reforms in the right direction; determining whether these reforms are "enough" is a very different matter, and one that requires the attention of the research and policy community.

Sharing the Burden of the International Debt Crisis

By STANLEY FISCHER*

Muddling through was the right strategy to handle the international debt crisis in 1982. Over the next four years, the debtor countries performed miracles on current account as they made massive interest payments amounting frequently to as much as 6 percent of GNP, with very little inflow of funds. But the price in terms of growth has been heavy, the debt problem continues to bedevil many of those countries, and the time for debt relief has arrived.[1]

The initial burden of the debt crisis was borne largely by the debtor countries. In addition, bank shareholders have paid the price of lower stock values for their assets. But because there has been no formal debt relief, debtor-country wage earners bear the burden of unnecessarily low real incomes. Formal debt relief will entail little further burden on bank shareholders, while substantially reducing the burden on wage earners in debtor countries. This should be possible without increasing burdens on taxpayers in the industrialized countries.

I. Background

There are two major debt crises—the crisis of Africa and that of Latin America and the Philippines. Although the African debt crisis is more serious in human terms, its implications for the financial system are less serious, and I do not discuss it further.

The debt crisis had three causes: imprudent macroeconomic management and borrowing by the debtor countries; imprudent lending by the commercial banks; and the increase in the *ex ante* real interest rate. The rise in the real interest rate to about 6 percent by 1982 increased the real interest burden on borrowers sixfold and completely changed the nature of the debt problem. With a real interest rate of 1 percent, growing out of a debt overhang was easy; with a 6 percent real interest rate, few countries could realistically hope that growth would easily reduce the debt burden without significant current account improvements.

It is unlikely that the possibility and effects of an interest rate shock of that magnitude were taken into account when the original loan agreements were entered into.[2] In the first instance, such an increase in the *ex ante* rate over a protracted period had not been seen previously. Second, it is likely that the significance of the shift from fixed to floating rate lending had not been absorbed. In past debt crises when loans were made at fixed interest, real interest rates would rise with deflation. But once the price level stabilized, the real interest burden would be higher only to the extent of the proportional decline in the price level. And it remained possible that inflation would reduce the burden in the future. In this crisis, the real interest rate has risen and stayed high for five years, and shows little sign of falling soon. Now inflation brings no automatic debt relief.

The initial response to the debt crisis in 1982 was to work out debt reschedulings with IMF approval certifying the debtors'. macroeconomic policies. That was the appropriate strategy. Something—renewed world growth, a recovery of primary product prices, or a decline in the real interest

*Department of Economics, MIT, Cambridge, MA 02139, and Research Associate, NBER.

[1] Rudiger Dornbusch (1986) and Paul Krugman (1986) both discuss the case for debt relief, as does Jeffrey Sachs (1986).

[2] The *Brookings Papers on Economic Activity* (*BPEA*), representative of thinking by well-informed economists, contains only three papers on the international debt between 1977 and 1982. Robert Solomon (1977) contains no suggestion that future interest rate developments might affect the stability of the debt. In the discussion following the paper, the real interest rate receives mention only in the last sentence. In papers in 1981, Solomon and Sachs mention that the debt problem could become serious if the real interest rate failed to come down—but place no emphasis on this possibility.

TABLE 1—DEBT DATA, DEVELOPING COUNTRIES, WESTERN HEMISPHERE

	1978	1982	1986
Total ($b)	155.8	333.0	382.5
Debt/Exports (%)	217.0	273.1	333.1
Debt Service /Exports (%)	38.2	50.6	46.0
Debt/GNP (%)	31.8	43.5	47.0

Source: World Economic Outlook, International Monetary Fund, October 1986.
Note: Total debt outstanding for all capital-importing developing countries was $399b. in 1978, $763b. in 1982, and $973b. in 1986.

rate—might turn up while the borrowers made much-needed changes in macroeconomic policy. With the rapid U.S. recovery in 1983–84 and the impressive turnarounds in trade account by the debtors, proponents found justification for their positions.[3] To be sure, the continued high real interest rate was a source of worry, but, in 1985, Gramm-Rudman-Hollings promised improvement on that front.

Developments in the debtor countries were less encouraging. There was indeed an extraordinary turnaround in trade account: in Brazil and Mexico an improvement of the order of 6–7 percent of GNP, compared with the worsening of the U.S. trade account of 3 percent of GNP that is proving so difficult to reverse. But the improvement in debtor country trade accounts was a result of import compression, through real devaluation and restrictive aggregate demand policies. Export volume had risen since 1980, but low dollar prices kept the dollar value of exports virtually unchanged. Investment fell by 5–7 percent of GNP between 1981 and 1985 for the major debtors.

Per capita real GNP was down, as much as 20 percent in some of the debtor countries.[4] The debt-to-exports ratio was not falling despite current account improvements and the rising volume of exports. More concretely, the burden of the debt took the form of transfers from the developing to the developed countries as capital inflows slowed. Whereas in 1981, Latin America and the Caribbean had a net inflow of resources of 2 percent of GNP, in 1984 and 1985 that area was transferring nearly 4 percent of GNP abroad.[5]

II. The Role of the Banks

By 1984 the commercial banks had become battle weary. Interest receipts from debtor countries exceeded new loans in 1984, 1985, and 1986. The U.S. banks actually reduced their exposure in Latin America in 1985. The banks, the main beneficiaries of the efforts of the IMF and the U.S. authorities' attempts to maintain continued debt servicing, participated in IMF packages with increasing reluctance. The Baker Plan in October 1985 spelled out the conditions for continuing cooperation among the U.S. government, the multilateral lending institutions, and the banks, but still did not succeed in bringing new commercial bank financing of even the modest proportions included in the plan into the developing countries.

Fear of financial collapse in the United States was one of the main motivations for the original approach to the debt crisis. In 1982, the nine large money center banks had over 250 percent of their capital in loans to LDCs; the proportion for all U.S. banks taken together was above 150 percent. By mid-1986, the nine money center banks had sufficient equity and reserves to withstand even the complete loss of Latin American assets (Table 2). The European and Japanese banks, which built up loan loss reserves more rapidly, were even better placed than U.S. banks to withstand losses on LDC debt.

Although banks have increased their loss provisions, they continue to carry LDC debt

[3] The best-known scenario is that of William Cline (1984).

[4] Argentina and Venezuela suffered the largest income declines.

[5] The net transfer is calculated as the current account deficit (representing net capital inflows) minus net investment income, both taken from *World Economic Outlook* (October 1986, p. 77).

TABLE 2—U.S. BANK EXPOSURE TO DEVELOPING COUNTRIES (June 1986)

	Nine Money Center Banks	All U.S. Banks
To LDCs:		
$b	73.9	112.1
% of Capital	167.2	101.3
% of Assets	11.7	7.3
To Latin America and Caribbean:		
$b	43.2	68.2
% of Capital	97.7	61.6
% of Assets	6.9	4.4

Source: Federal Reserve Board of Governors, Statistical Release E16 (126), October 17, 1986. "Capital" consists of equity, subordinated debentures, and reserves for loan losses.

TABLE 3—SECONDARY MARKET LDC DEBT PRICES, NOVEMBER 14, 1986 (Cents/$)

Country	Price	Country	Price
Argentina	65.7	Mexico	57.2
Bolivia	7.5	Philippines	73.0
Brazil	75.5	Turkey	98.3
Chile	68.0	Venezuela	74.5

Source: Salomon Brothers, Inc.
Note: Price is average of bid and offer prices.

at face value. The active market in such debt prices it at a significant discount. Sample prices are listed in Table 3. Equity values for the banks are consistent with the prices of debt in the secondary market.[6]

III. Proposed Solutions

Reform proposals can be distinguished according to whether they propose extending the effective maturity of the debt, changing the nature of claims on the LDCs, whether they offer genuine debt relief, and what sort of conditionality they impose (Paul Krugman).

The present strategy and the Baker Plan both deal with the debt crisis by extending the effective maturity of the debt. Interest capitalization would do the same. Any method that reduces the current flow of resources from the debtor countries will help them grow in the short run. But further lending promises little in the way of a lasting solution to the debt problem. So long as real interest rates remain around 6 percent, debtor countries will have great difficulty growing out of their debt problem.

Debt-equity swaps, increased direct investment, proposals to sell shares in the export earnings of the debtors, and indexation of payments to export prices, would all change the form of the debt. None necessarily changes the present value of the debt, though their risk characteristics may make these forms of debt more attractive to the debtors. The recent Mexican agreement takes the IMF part way down the road of providing cash flows that respond to the risks facing the developing countries.

The debt crisis has made it entirely clear that floating rate debt is a poor way of financing a country's development. Innovative methods of financing all give promise that international capital flows can resume without producing the danger of another debt crisis. However, they are being introduced too slowly to resolve the current crisis.

Conditionality is explicit in the Baker Plan and in the present strategy. As IMF conditionality has progressed over the past four years to the sophistication of the recent Mexican package, it has become a viable means of providing useful external constraints on domestic policymakers. Conditionality will be needed if debt relief is granted, both to ensure that the relief is not wasted, and to prevent relief being an entirely pleasant experience.

IV. Debt Relief

There are now two choices. Either the piecemeal approach continues, or there is some form of debt relief. The current approach is certainly more imaginative—provided the Mexican package is not the last of its kind—than that of 1982. But the banks are increasingly reluctant to participate.

The argument that the slow approach is the right one points to the successes of the

[6] In a study based on 1983 data, Sachs and Steven Kyle (1984) suggested Latin debt was then being carried at about 80 cents on the dollar.

last four years—there was no financial collapse in the United States, there has been no explicit debt default, Latin America is moving towards instead of away from democracy. Maybe something will still come up; perhaps a U.S. tax increase that takes the real interest rate down by 2–3 percent, perhaps an improvement in the trade climate, perhaps a growth recovery.

Perhaps liberalization of debtor economies will solve the debt problem, as increasing confidence draws flight capital home. High interest rates might attract flight capital home, just as they may entice other capital from abroad. But the interest rates needed to bring home flight capital will not restore investment. Indeed, the notion that flight capital should come home is not consistent with general liberalization, for it is very likely that optimally diversified portfolios for residents of developing countries contain more industrialized country assets than they do at present—even including flight capital. Liberalization is likely to lead to more, not less, capital flight.

The case for debt relief is not that the present evolution cannot continue, but that it should not continue. For four years, the debtor countries have paid the price of low GNP growth and significant falls in real wages as they have made transfers to service the debt. Protectionist pressures in the industrialized counties have made the transfer process more difficult. The transfers have been made at a real interest rate that was almost certainly not envisaged when the debt was incurred.

Economic theory has little to say about the appropriate procedures to follow when unanticipated events happen.[7] Formally, the loan contracts do not give the borrower the right to reopen negotiations, nor is there any procedure for establishing whether a particular set of circumstances might reasonably have been anticipated when the contracts were entered into.

Presumably any potentially Pareto-improving changes in debt contracts have already been made. What remains are changes that improve the lot of one party to the contract at the expense of the other. Thomas Walde notes that there is a strong legal presumption in favor of lenders, who gave up real resources in exchange for promises. The IMF and developed country governments have certainly taken the attitude that the debt contracts should continue to be honored, no matter what the burden on the borrowers.

One argument for maintaining current debt contracts is distributional—that the lenders deserve to receive the payments due them. That is a hard argument to support in the present crisis. The lenders had no reason to expect such payments, they too made mistakes, and they are not obviously more deserving than the borrowers. Further, the borrowers have paid a high price for whatever mistakes they made in the past.

The market, reflected in Table 3, has already concluded that the lenders will not receive their claims in full. Shareholders have taken their losses. But the lenders hang on to their claims in the hope of capital gains in the event the borrowers pay in full. Millions of residents of developing countries are being kept at low levels of income for the sake of possible capital gains for bank shareholders. There is no distributional case for the current debt strategy. There is a strong distributional case for debt relief.

Even so, there might be an efficiency argument against relief. It has been contended from the beginning of the debt crisis that only by maintaining existing debts can the existing international capital markets be maintained. It is impossible to be sure, but the evidence from history is strong that default or relief is not the end of the capital markets. Indeed, debt relief that promises to put an end to the uncertainty of the current situation would likely promote future capital flows—albeit, and to the good, in forms other than floating rate debt. It is difficult indeed to believe that international capital

[7] Thomas Walde (1986) discusses legal precedents in domestic and international agreements, including force majeure and change of circumstances. The most relevant cases appear to be the Westinghouse and Alcoa vs. Essex cases where price shocks resulted in courts excusing nonperformance.

will fail to flow to a country offering good rates of return.

Moral hazard is another argument against debt relief. Why grant relief to the badly behaved when the well-behaved, such as Turkey, have made a more serious and successful effort to adjust? Isn't this an invitation to countries in trouble in the future to default? The answers here are simple. No country will get debt relief in the future without going through a protracted period of uncertainty and adjustment. What borrower will in the future become overextended to have the privilege of following in the footsteps of an Argentina or Mexico or Brazil over the past four year? Further, the proposal is for debt relief administered from the center, not for default by the borrowers.

What form should relief take? Relief should only be available in the context of structural adjustment programs adopted by the countries in cooperation with the IMF and, depending on how rapidly it adapts, the World Bank. There would be no general forgiveness of debt, rather in each new negotiation interest and principal payments to commercial lenders would be reduced to 65 percent of the contractual value contingent on a structural adjustment program being agreed with the IMF. If the commercial lenders found such terms inappropriate, they would forego the help of the IMF and their governments in extracting resources from the debtors, and negotiate on their own.

The IMF agreements would be for comprehensive growth-oriented adjustment programs, encouraging investment in both government and private sectors. Other desirable structural adjustments, along the lines of freeing up markets and privatization of sectors that have no inherent government connection, such as nightclubs, hotels, and grocery chains, could be included. So too would the opening up of the economy to imports, and the removal of standard developing country distortions in the form of subsidies.

What would such a scheme cost? The total outstanding debt to financial institutions and other private creditors of all countries with recent debt-servicing difficulties was $336 billion in 1986. Interest payments on this amount appear to be about $28 billion.[8] The annual reduction in the interest bill would be close to $10 billion. For example, Brazil and Mexico would save about $2.5 billion annually, the Philippines and Chile about $600 million. The numbers are between 1 and 3 percent of GNP, enough to make a difference, but not enough to let the debtor countries off the hook entirely.

What would be the consequences for the banks? They would take a loss, but there should be no mass bankruptcies. The losses would be recorded gradually as countries negotiated agreements with the IMF, rather than on announcement of the plan.

Without debt relief, the debt crisis promises to drag on for decades, slowing growth in the developing countries, sapping the energies of policymakers, and tieing up the multilateral lending agencies in endless crisis negotiations. With sensible debt relief, countries and the multilateral institutions can begin to worry about growth-oriented development policies. If the debt relief does not come by agreement, then debtor countries would have to consider taking the first step.

[8] Based on data in *World Economic Outlook*, October 1986.

REFERENCES

Cline, W. R., *International Debt*, Washington: Institute for International Economics, 1984.

Dornbusch, R., "The Debt Problem and Some Solutions," mimeo., Department of Economics, MIT, November 1986.

Krugman, P., "Prospects for International Debt Reform," report to the Group of Twenty Four prepared for UNCTAD, January 1986.

Sachs, J., "The Current Account and Macroeconomic Adjustment in the 1970s," *Brookings Papers on Economic Activity*, 1:1981, 201–68.

______, "Managing the LDC Debt Crisis," *Brookings Papers on Economic Activity*, 2:1986, 397–431.

_____ **and Kyle, S.**, "Developing Country Debt and the Market Value of Large Commercial Banks," NBER Working Paper No 1470, 1984.

Solomon, R., "A Perspective on the Debt of Developing Countries," *Brookings Papers on Economic Activity*, 2:1977, 479–502.

_____, "The Debt of Developing Countries: Another Look," *Brookings Papers on Economic Activity*, 2:1981, 593–606.

Walde, T., "Sanctity of Debt and Insolvent Countries, Defenses of Debtors in International Loan Agreements," mimeo., UN, New York, May 1986.

The Debt Crisis and the Future of International Bank Lending

By H. ROBERT HELLER*

As we approach 1987, the Gordian knot of the international debt problem seems to be tightening again. Rescheduling agreements and requests for new money by several important countries are on the agenda. It will take farsighted and strategic thinking on behalf of all participants to overcome the mounting problems confronting us. Nevertheless, I firmly believe that the problems will continue to be manageable if all participants focus on their long-term interest in coming to a satisfactory solution. It is within that general framework that I would like to approach the topic of this paper.

The role of the banks in the debt crisis cannot be seen in isolation. Important interdependences must be considered, including the responsibilities of the debtor countries, the industrialized countries, and the international agencies.

I. The Causes of the Debt Crisis and Initial Management Success

During the 1970's, commercial banks became the principal source of external finance to the developing countries. The volume of new bank-related financial flows surpassed by far the volume of official government lending, financing provided by the multinational agencies, and foreign direct investment.

The global recession of the early 1980's and the associated fall in commodity prices produced a sharp curtailment of the earnings of the developing countries. At the same time, real and nominal interest rates surged, straining the financial resources of the LDCs until they were no longer able to fulfill their financial commitments, thereby triggering the international debt crisis of 1982.

The initial management of the debt crisis has been generally satisfactory. Immediate liquidity assistance was provided by the central banks of the industrialized countries under the auspices of the Bank for International Settlements. Subsequently, the International Monetary Fund assumed a central role by assisting in the implementation of financial adjustment programs and providing $35 billion in financial assistance between mid-1982 and mid-1986. Allowing for $12 billion in repayments (albeit not necessarily by the same countries), net funds provided amounted to $23 billion.

The World Bank also made significant contributions by disbursing over $32 billion in new loans through the IBRD over the period from mid-1982 to mid-1986. Allowing for $11 billion in loan repayments, the net new money provided by the IBRD amounted to $21 billion. Adding the concessional IDA credits further enhances the role of the World Bank. During the same 1982–86 period, IDA disbursed $10.8 billion. Repayments amounted to less than half a billion, resulting in a net new IDA money flow of $10.4 billion.

Over the same period, commercial banks increased their net LDC exposure by $45 billion. Debt reschedulings were agreed upon between the banks and all the major debtor countries. Banks, and in particular U.S. banks, also increased their capital rapidly during the period, thereby reducing the ratio of outstanding LDC debt to their own capital and enhancing their capacity to withstand possible adverse developments.

Many of the developing countries implemented tough financial adjustment programs that resulted by 1985 in a drop in imports by $82 billion or 14 percent below the 1981 level. Of course, lower export prices also played a role in this drop of export revenue, but the end result was just as painful for the developing countries.

At the end of 1986, we can look back upon a record of considerable accomplish-

*Member, Board of Governors of the Federal Reserve System, 20th and Constitution Ave., NW., Washington, D.C. 20551.

ments both for the debtors and the creditors. Key to that progress has been the cooperative attitude between borrowers and lenders, facilitated by the good offices and resources provided by governments and international agencies.

II. A Strategic Perspective on International Bank Lending

If the international debt program is to be solved in a cooperative fashion with the continued active participation of the commercial banks, it is important to investigate whether the long-term interests of the banks are served by such a course of action.

Similarly, from the viewpoint of the debtor countries, the key question is whether the countries wish to maintain a long-term relationship with foreign commercial banks. The costs associated with debt repudiation became abundantly clear in the case of Cuba, and no country that wishes to remain part of the international commercial and financial system will consider this a viable alternative. In more recent history, Peru's unilateral decision to limit its debt service payments has not put that country on a path to economic health and prosperity.

The formation of valid expectations about the future of international bank lending requires an understanding of what motivates a bank. In that sense, the future of international bank lending to developing countries should be seen as a long-term strategic decision and not merely a credit judgment about a particular borrower at a given moment in time or the expected rate of return on a specific loan.

If banks reduce and eventually eliminate their international activities either by engaging in a slow retreat or by divesting the international division, they engage in a strategic retreat that will not be easily reversed within a decade or more. The central issue is whether it is in the banks' own long-term interest to maintain a business relationship with the developing countries.

The business interests of commercial banks are highly complex, and it is clear that not all aspects can be fully considered within the limited space available. It should also be borne in mind that not all banks are similarly situated, and that business judgements by the senior managers may well differ.

With these caveats in mind, it is useful to consider three broad areas of foreign banking activity in arriving at an assessment of the future prospects of international bank lending: the international short-term transaction business, medium- and long-term lending, and direct investment activities in the international banking area.

The first segment of international banking activities comprises payment and transaction-oriented services. Examples include the execution of international payments, foreign exchange services, letters of credit, trade finance, traveler's checks, and credit cards. Many of these transactions are self-liquidating and arise in connection with the payment for international goods and service flows. The total amount of cross-border exposure that will be generated by these transactions may be quite small. These services are generally provided by a small number of banks with an extensive international presence: the international network banks. Only banks with an extensive international network are able to compete in this market effectively as principal players. Other banks partake in these transactions on a one-off basis through correspondent banks.

Historically, the largest North American and European banks have dominated this market, although in recent years Japanese banks have made significant efforts to become major players as well. The global total of banks in this category is probably less than 20, and due to the high costs of entry, there is little prospect that the number of banks in this category will increase significantly. Conversely, the banks active in this market see their franchise as a valuable possession and are not likely to give up this area of activity.

However, banks may well question how many countries need to be included in such a global network strategy. The potential market size, and thereby profit potential, drops rapidly after the largest countries are accounted for. The smaller countries may therefore find it increasingly difficult to retain a large number of banks that are willing to incur the substantial up-front costs of maintaining the necessary offices and overhead expenses. On the other hand, there is a strong self-interest on behalf of the banks to remain active in the large countries.

The second market segment involves the medium- and long-term financing of projects and balance of payments imbalances. By their very nature, these activities involve the creation of relatively large and persistent cross-border credit exposures.

In years past, innovative financial techniques, such as syndicated Eurocurrency credits, made it possible for many banks to participate in these lending activities. The lead banks were able to spread the associated risk among a large group of smaller banks that participated in the syndications. In addition, the large banks benefited from the income derived from the management and syndication fees.

There is a broad tendency in financial markets toward the securitization of lending. International markets will not be exempt from that trend. International loan syndications are therefore likely to be replaced by the securitization technique. It also stands to reason that securitization is better suited for traditional project lending with an identifiable stream of repayments rather than for balance-of-payments lending to governments, where the availability of funds for repayment often depends on the implementation of appropriate policy measures.

In order to gain access to the international security markets, the credit quality of the borrowers must be unquestioned. It is therefore in the interest of both the debtor countries and the creditor banks to do everything in their power to enhance the creditworthiness of the debtor nations, so that the security markets can eventually be used to offer a broader access to world financial resources to the developing countries.

Debt-to-equity swaps offer one tool to ease the fixed payments commitments of the debtor countries. Several commercial banks have pioneered debt-to-equity swaps and many more promising opportunities exist in this field. For instance, the possibility of "mutual funds" that could be used by smaller banks to pool their equity investments is now being explored. But even under the best of circumstances debt-to-equity swaps will make only a marginal contribution towards the resolution of the entire problem. This assumes the full cooperation of the debtor countries in liberalizing their regulations towards foreign direct investment.

Difficult regulatory and supervisory issues may arise in connection with debt-to-equity swaps. I believe that it will be possible to overcome these problems, so that debt-to-equity swaps can make a useful contribution by reducing the magnitude of the fixed payment commitments of the debtor countries.

Both from the viewpoint of the banks and of the countries involved, it is desirable that debt-to-equity swaps involve more than a mere liquefaction of the initial loan at a discount. From the viewpoint of the country, it is desirable that the debt-to-equity conversion constitutes a net addition to the foreign direct investment inflows. From the viewpoint of the bank, it is desirable that the bank be in a position to partake in the potential profits that the equity investment affords to the holder. Merely selling a loan at a discount to a potential foreign investor does not fulfill that objective.

New investments in the financial service sector in the various countries are particularly well suited to fulfill the dual objective. The country's consumers and business establishments will gain by increasing competition in the financial service sector, the bank gains new business opportunities in a field of its own expertise, and the country and the bank forge a long-term strategic relationship.

This brings up the third strategic business relationship between banks and foreign countries: the local presence of foreign banks via branches and/or subsidiaries. Banks in this category will have a continuing involvement with the domestic economy of the country and engage in a wide variety of financial activities. It goes without saying that the deepening of such a relationship is very much in the interest of both the banks and the countries. Not only will the banks involvement with the local economy enhance their interest in the economic health of the country, but the country may gain by having access to new and varied sources of financial services. It therefore stands to reason that the rapid liberalization of financial markets in the developing countries by opening the markets to full and equal participation by foreign banks on a "national treatment" basis offers a promising avenue that will increase the long-term congruence of interests among debtor countries and their creditor banks.

The issue whether some of the LDC loans should be written down or "marked to market" has been discussed frequently in that connection. My personal view is that there exists good reason to treat a loan on the books of a bank at face value as long as there is a reasonable expectation that the debt service payments will be made and that the loan will eventually be repaid. The bank may enhance the likelihood of this event by participating in restructuring and new lending arrangements. Such actions may therefore lend credence to a bank's expectation of eventual repayment. In contrast, banks that refuse to partake in refinancing activities may thereby signal their own doubt as to the future collectability of the loan and may wish to make appropriate value adjustments.

III. The Responsibility of the Industrialized Countries

The international debt problem cannot be solved on a bilateral basis between the banks and the debtor countries. To a large extent, these bilateral negotiations involve arguing about who should make the bigger sacrifice: the debtor or the creditor. With the exception of the new investment opportunities in the financial sector discussed above, the argument revolves around redistributive questions rather than an enlargement of the total amount of economic and financial resources available to both parties.

Ultimately, the international debt crisis can be overcome only by enlarging the economic pie through economic growth and increased exports by the debtor countries. This involves not only a concerted effort by the debtor countries to increase their exports, but also continued access to markets in the industrialized countries. This latter condition is absolutely essential in order to allow the debtor countries to earn the foreign exchange needed to service the debt.

Total exports by the developing countries amounted to $595 billion in 1981, the last year before the debt crisis, but fell to $508 billion by 1985. While much of that shrinkage in exports was due to reduced trade among the developing countries themselves, the industrialized countries reduced their purchases from the developing world by over $50 billion. The LDC exports to the United States remained essentially constant with $102 billion in 1981 and $101 billion in 1985. However, LDC exports to Japan dropped from $79 to $67 billion over the same period. Other countries with sharply declining imports from the LDCs include Australia, Belgium, Canada, France, Germany, the Netherlands, Sweden, Switzerland, and the United Kingdom.

It is true that much of that change in trade was accounted for by lower petroleum prices. All the more important is it that the industrialized countries that benefited the most from the fall in petroleum prices and commodity prices utilized the funds saved to purchase more products from the developing countries.

Much of the record trade and current account surpluses enjoyed by countries like Germany and Japan are not due to the country's own greater export efforts, but due to reduced purchases from the LDCs. Thus it is clear that there exists much room to expand imports from the developing countries. Without access to growing markets, it will be difficult for the debtor countries to earn the foreign exchange needed to make the debt service payments. The balance-of-payments surplus countries are in a unique position to make a contribution toward overcoming the international debt service problems and increased international financial balance and stability. The proper way to reestablish the creditworthiness of the developing countries is through more trade and not through debt relief.

IV. The Role of the International Organizations

The IMF and the World Bank must continue to play a leadership role not only by providing continued financial assistance, but also by assisting in the implementation of market-focused adjustment programs in the debtor countries. Strong pressure on the surplus countries to allow the necessary adjustments in their external imbalances is also called for. The record profits enjoyed by the World Bank place it in a unique position to use some of these profits to grant loans at lower interest rates. In addition, it may utilize its seldom used authority to guarantee commercial bank loans in order to serve as a catalyst for more private lending. By guaranteeing, say, 50 percent of a commercial bank loan, the World Bank may effectively double the leverage of its own scarce capital.

V. Conclusion

There is no doubt that international debt service problems will remain an important concern in the years to come. There is also no doubt that it is in the long-term strategic interests of the creditor banks and the debtor countries to cooperate in overcoming the current difficulties.

Banks will have to continue to make temporary sacrifices if they want to see the long-term quality of their assets improve and if they wish to maintain their strategic interests as international financial institutions. Some banks may see their own strategic interests better served by equity investments rather than a continuation of debt holdings.

The debtor countries can enhance the long-term strategic interests of the banks by opening up their financial markets and granting new franchises. A continuation of the adjustment effort that focuses on strengthened export performance, rather than arbitrary import curtailment, is also called for.

The industrial countries with strong current account positions are now called upon to lead the world toward a more open and growth-oriented economic and financial system that will not only benefit the developing countries in need of export markets, but also make their own citizens better off by providing them with a wider and cheaper variety of products.

The IMF and the World Bank should continue to play the key role envisioned in their respective charters as the providers of balance-of-payments assistance and development finance. In addition, they must serve as a forum for the international policy discussions on appropriate adjustment measures to be implemented by surplus and deficit countries alike.

Understanding Rent Dissipation: On the Use of Game Theory in Industrial Organization

By DREW FUDENBERG AND JEAN TIROLE*

Game theory has had a deep impact on the theory of industrial organization, in a similar (but less controversial) way as the rational expectations revolution in macroeconomics. The reason it has been embraced by a majority of researchers in the field is that it imposes some discipline on theoretical thinking. It forces economists to clearly specify the strategic variables, their timing, and the information structure faced by firms. As is often the case in economics, the researcher learns as much from constructing the model (the "extensive form") as from solving it because in constructing the model one is led to examine its realism. (Is the timing of entry plausible? Which variables are costly to change in the short run? Can firms observe their rivals' prices, capacities, or technologies in the industry under consideration? Etc.)

A drawback of the use of game theory is the freedom left to the modeler when choosing the extensive form. Therefore, economists have long been tempted to use so-called reduced forms, which try to summarize the more complicated real world game, as for example in the literature on conjectural variations, including the kinked demand curve story. This approach is attractive, but has several problems. The obvious one is that the modeler can only be sure that the reduced form yields the solution of a full-fledged model if he has explicitly solved the model. Also, reduced forms are most natural for the description of steady states, and are thus ill-suited to describe battles for market shares (like price wars, predation, entry and exit), or to study the adjustment paths to outside shocks or government intervention. (Reduced forms are not robust to structural changes.) While the reduced-form approach is simpler, and so more amenable to applications, we believe that the focus on "primitives" implied by the extensive-form approach allows a clearer assessment of the model. Furthermore, the diversity of "reasonable" extensive forms may to some extent reflect the wealth of strategic situations in industries.

This paper illustrates how game theory sheds light on a particular issue—the hypothesis of monopoly rent dissipation advanced by Richard Posner (1975).

Because rents accrue from a monopoly (or oligopoly) position, rent-seeking behavior is an important phenomenon in industrial organization. Following the rent-seeking literature, Posner argues that firms in general engage in a contest for monopoly power and suggests that all monopoly profits may be added to the usual deadweight loss triangle (associated with monopoly pricing) in the social costs of monopoly. The main postulates behind this analysis are 1) *The "rent dissipation postulate"*: the total expenditure by firms to obtain the monopoly profit is equal to the level of this profit, and 2) *The "wastefulness postulate"*: This expenditure has no socially valuable by-products.[1]

†*Discussants*: Timothy F. Bresnahan, Stanford University; Richard L. Schmalensee, Massachusetts Institute of Technology; Robert D. Willig, Princeton University.

*Departments of Economics, University of California, Berkeley, CA, 94720, and, Massachusetts Institute of Technology, Cambridge, MA, 02139, respectively. We thank Timothy Bresnahan, Richard Schmalensee, and Robert Willig for helpful comments.

[1] See Franklin Fisher (1985) for a useful discussion of Posner's argument.

If there are a great many potential contenders for the rents, all of whom are exactly as efficient, then one expects that "free entry" will ensure that the rent dissipation postulate is satisfied. Even here, the wastefulness postulate is not always appropriate, as we will see. Moreover, there are many situations in which neither postulate is appropriate. Extreme incumbency advantages may allow established firms to blockade entry, and appropriate the entire rent; in some cases, like that discussed in Section III, this can occur even if the incumbency advantage seems "small." We will analyze three different forms of rent-seeking behavior, assuming that firms are symmetric or nearly so. We argue that the game-theoretic approach helps to illuminate the key features of these contexts, and to explain the nature of rent dissipation in each.

I. Natural Monopoly

A. *Contestability*

One of the most provocative contributions in industrial organization in recent years has been the theory of contestability (see William Baumol, John Panzar, and Robert Willig, 1982), which develops the idea that potential competition may be as effective as real competition. This idea is particularly important for industries in which increasing returns to scale (for example, fixed costs) limit the number of firms in the market (natural monopoly or oligopoly). Suppose that all firms in an industry have the same cost function. An industry configuration is a list of outputs for each firm (zero for the "entrants," positive for the "incumbents"). A configuration is *feasible* is there is a price p at which supply equals demand, and each firm at least breaks even; it is *sustainable* if no entrant (a nonproducing firm) could find a price p^e under p and an output not exceeding the demand at p^e such that it makes a strictly positive profit, given that the incumbent firms stick to price p. That is, in a "perfectly contestable market," entrants have the same technology as incumbents and take the latters' price as given.

Let us illustrate the idea in a simple one-good example.[2] Suppose that the common cost function is $C(q) = f + cq$ if $q > 0$, $C(0) = 0$, where f is a fixed cost. Let $\tilde{\Pi}^m = \max_q \{qP(q) - cq\}$ denote the variable monopoly profit, where $P(\cdot)$ is the inverse demand function. Suppose $f < \tilde{\Pi}^m$. The unique sustainable price is given by $(p^c - c)D(p^c) = f$. Any price under p^c yields a negative profit (price under average cost) and any price above p^c is undercut by an entrant. In this example, contestability predicts:

Conclusion 1: *There is a unique firm in the industry* (*technological efficiency*).

Conclusion 2: *This firm makes zero profit.*

Conclusion 3: *Average cost pricing prevails. Furthermore, the allocation is socially efficient given the constraint that regulators do not use subsidies.*

Conclusion 2 is in accordance with the rent dissipation postulate of the rent-seeking literature. Conclusion 3 is at odds with the wastefulness postulate. Rents are dissipated in a socially efficient manner by competitive pricing.

The natural question is which situation the contestability axioms depict. Unlike the conjectural variations approach, contestability can be given a firm game-theoretic foundation. The price p^c is the Nash equilibrium outcome of a game in which firms choose their prices first and then decide whether to produce. The proponents of contestability have developed a similar justification, based on the idea of "hit-and-run entry." In this model, if the incumbent's price exceeds p^c, an entrant will enter at a lower price and make a profit before the incumbent has time to lower its own price in response. The hit-and-run model has been criticized on the grounds that prices seem to adjust faster than quantities, even in industries with "capital on wheels" like the airline industry. The debate that ensued is, we believe, a demonstration of game theory's usefulness. It highlights important aspects of the timing

[2] This example does not do full justice to the theory because many interesting aspects of contestability relate to multiproduct monopolies and cross subsidization.

and commitment structures of strategic competition in industrial markets.

We next present two alternative, more dynamic approaches to natural monopoly, one at odds with, and the other possibly more sympathetic to, the contestability approach. Both reverse the logic of contestability by assuming that price adjustments are faster (actually instantaneous) than quantity adjustments.

B. *War of Attrition*

Consider the previous model in continuous time with interest rate r. There are two firms, and price adjustments are instantaneous. If the two firms are in the market, price equals marginal cost c, and both firms lose f. If only one firm is in the market, the price is equal to the monopoly price, and this firm makes profit $\tilde{\Pi}^m - f > 0$. Both firms are in at date 0. Each firm chooses a date at which to exit (conditional on the other firm still being in the market at that date). We claim that the following symmetric behavior forms a (perfect) equilibrium: If both are still in at date t, each firm drops out with probability $rf/(\tilde{\Pi}^m - f)\,dt \equiv x\,dt$ between t and $t + dt$ and never returns; the remaining firm remains in forever. To check, note that each firm's expected profit from any date t on, given that both firms are still in, is equal to zero, since the firm is willing to drop out immediately. By waiting, it loses $f\,dt$ and gains the present discounted value of monopoly profits $(\tilde{\Pi}^m - f)/r$ with probability $x\,dt$.

A few remarks about this equilibrium. First, it is consistent with free entry and exit (in that it is still an equilibrium under costless exit and reentry). Second, it is not unique; however, if one allows a sufficiently large symmetric uncertainty about the rival's fixed (or opportunity) cost (here, the fixed cost is common knowledge), the symmetric equilibrium is the unique equilibrium, as shown in our 1986 paper. See that paper for references to the literature.

The war of attrition equilibrium yields

Conclusion 1′: *There are two firms in the industry for a* (*random*) *length of time and then one* (*technological inefficiency*).

Conclusion 2′: *Firms earn no ex ante rents* (*but may have ex post profits*).

Conclusion 3′: *The price is first competitive and then is equal to the monopoly price. The allocation is not "constrained efficient"* (*and welfare is lower than under contestability, as is easily checked*).

This war of attrition, like contestability, satisfies the rent dissipation postulate. However, there is some wasteful dissipation, although the waste is not complete. Because consumers pay the monopoly price only after the shake-out period, welfare is higher than Posner would predict. So this outcome is intermediate between the solutions suggested by Baumol-Panzar-Willig and Posner from the point of view of the wastefulness postulate.

C. *Short-Run Commitments*

Let us now assume that capital is sunk in the short run. The first theory of short-run commitments was developed by B. Curtis Eaton and Richard Lipsey (1981). One unit of capital is necessary for production, and gives access to constant marginal cost c. Capital has cost f per unit of time and has durability H. If only one firm is active (has capital), its profit gross of capital cost is $\tilde{\Pi}^m \varepsilon(f, 2f)$. If two firms operate (have capital), they make zero gross profit. The firm's sole decision is when to build a new unit of capital. One firm gets to invest first at time 0. Eaton and Lipsey construct the following (symmetric) strategies. The incumbent firm (the one with a plant) always builds a new plant $\Delta(< H/2)$ years before its current plant depreciates. The other firm invests in a plant if the incumbent has only one plant and this plant is more than $H - \Delta$ years old. In equilibrium, the length Δ is chosen such that at $H - \Delta$ the entrant is indifferent between entering and not entering. If he does not enter, the incumbent stays a monopoly forever, and the entrant makes no profit. If he enters, the entrant makes profit $(-f)$ for Δ years (because the incumbent is still committed to his old plant), and then enjoys the monopoly profit forever.

Let

$$V=\frac{\tilde{\Pi}^m}{r}-\frac{f}{r}\left(\frac{1-e^{-rH}}{1-e^{-r(H-\Delta)}}\right)$$

denote the present discounted profit of the incumbent from date zero on. If he enters, the entrant gets V minus the loss in monopoly profit during Δ years. That is;

$$(1)\qquad V-\tilde{\Pi}^m((1-e^{-r\Delta})/r)=0.$$

or

$$(2)\quad \tilde{\Pi}^m/f=(1-e^{-rH})/(e^{-r\Delta}-e^{-rH}).$$

It is easily checked that, given our assumptions, (2) implies that $\Delta < H/2$. We are particularly interested in what happens as length of commitment H tends to zero. From (1), V tends to zero. And from (2), Δ/H tends to $1-(f/\tilde{\Pi}^m)$. The Eaton-Lipsey model thus yields for very short commitments:

Conclusion 1″: *There is only one firm in the industry at any point in time. (But the industry is not technologically efficient, because this firm has two units of capacity $(1-f/\tilde{\Pi}^m)$ percent of the time).*

Conclusion 2″: *Firms make no profit.*

Conclusion 3″: *The monopoly price prevails.*

This model yields Posner's postulates 1 and 2 exactly, as the monopoly rent is dissipated in a completely wasteful way. This should not be surprising—the *only possible* avenue for rent dissipation in this model is excess capital, which (by definition) has no social value.

Eric Maskin and Tirole (1986) point out that wastefulness is contingent on Eaton and Lipsey's assumption that the extra unit of capital cannot be used for production. They provide a different, discrete-time model of short-run commitments to capital in which firms can produce at zero cost up to their current capacities, which are locked in for two periods. Firms choose their capacities sequentially, and the horizon is infinite.[3] Let $\tilde{\Pi}(q_1,q_2)$ denote the per period profit of a firm with capacity q_1 when its rival has capacity q_2, gross of the per period fixed cost f (the usual assumptions on $\tilde{\Pi}$ are made, including a negative cross-partial derivative). Let $\delta=e^{-rT}$ denote the discount factor (T is the period length). As in Eaton-Lipsey, there exists a unique symmetric Markov perfect equilibrium. For sufficiently large δ, a firm chooses to enter if and only if its rival has capacity less than the entry-deterring capacity $\bar{q}$; if it enters, it accumulates capacity $\bar{q}$ itself. In equilibrium, $\bar{q}$ is such that the entrant is indifferent between entering and not entering:

$$(3)\quad \tilde{\Pi}(\bar{q},\bar{q})+\frac{\delta}{1-\delta}\tilde{\Pi}(\bar{q},0)=\frac{f}{1-\delta}.$$

(The entrant gets $\tilde{\Pi}(\bar{q},\bar{q})-f<0$ when entering and then becomes an "entry-deterring monopolist" forever.)

From (3), we see that for very short commitments, the per period monopoly profit $\tilde{\Pi}(\bar{q},0)-f$ converges to zero. To investigate the wastefulness postulate, we must wonder whether the monopolist's capacity $\bar{q}$ is equal to or exceeds this production. For this, assume that capacity installation has unit cost c_0 and fixed cost f. So, if q denotes production, $\tilde{\Pi}(\bar{q},0)-f=\max_{q\leq\bar{q}}\{qP(q)-(c_0\bar{q}+f)\}$. It is easily checked that if the installation cost is large, the monopolist produces to capacity: $q=\bar{q}$. We thus conclude in this case that for very small commitments, the outcome is exactly that predicted by contestability. Rent dissipation operates through price reduction, rather than through excess capacity.

II. The Adoption of New Technology

Now we turn to a model in which the "strategic" choice variable for the firms is the time to adopt a new technology. All firms know that due to technological progress, the cost of adopting the technology will fall over time. The model is set in continuous time, with interest rate r. Let $p(t)$ be the (perfectly foreseen) cost, in time-zero dollars, of adopting the technology at

[3] Alternatively, the model is equivalent to a continuous-time model in which firms choose capacities q which, as in Eaton-Lipsey, depreciate in a one-horse-shay manner, but according to a Poisson process (stochastic H).

time t: this is a sunk cost. We assume that $(p(t)\exp(rt))' < 0$, that $(p(t)\exp(rt))'' < 0$, that $p(0)$ is "large" (no one wants to adopt at zero) and that $\exp(rt)p(t) \to 0$ as $t \to \infty$, (so that if there are gains from adoption, then eventually adoption will occur).

We will consider two different forms of the rent that the new technology provides, and also two specifications of the information structure. As in our 1985 paper, we focus on the case of negligible information lags, where firms can observe and respond to their opponent's action without delay. Here, the firms may be tempted to adopt the technology "early" in order to delay or prevent adoption by an opponent. There is always an equilibrium in which rents are completely dissipated by "preemptive adoption." In our first example, this is the only (perfect) equilibrium. The second example shows that for a different specification of the "product market," there is also an equilibrium in which the firms can credibly arrange to postpone adoption until the time that is privately optimal, and thus retain some of the rents. Changing the information structure of the game yields dramatically different results. The "open-loop" information structure considered by Jennifer Reinganum (1981) corresponds to infinite observation lags, that is, the firms *cannot* observe their opponent's play. In this case, firms are not tempted to adopt preemptively, and rents are only partially dissipated.

A. *Imitation Deterring Innovation*

To illustrate our first outcome, consider a Bertrand duopoly, in which the two firms initially have constant unit cost c_0. Each firm makes no profit. Adopting the innovation reduces the unit cost to $c_1 < c_0$. Let $V = (c_0 - c_1)/r$. When only one firm has adopted, this firm makes Bertrand profit $(c_0 - c_1)$ per unit of time, assuming that c_0 does not exceed the monopoly price at cost c_1 ("nondrastic" innovation). Note that imitation never occurs in this extreme model, because Bertrand competition with identical costs yields zero-flow profit. The equilibrium time of adoption, t^c, is given by $V = p(t^c)e^{rt^c}$. If one firm, say firm one, planned to wait to adopt after t^c, then firm two would do better to adopt just slightly earlier. Thus, any proposed equilibrium adoption time $\tilde{t}$ with $\tilde{t} > t^c$ is vulnerable to preemptive adoption. This intuition was informally put forward by Partha Dasgupta and Joseph Stiglitz (1980). (The equilibrium turns out to require mixed strategies of a special kind, which we developed in our 1985 paper.)

The outcome satisfies Posner's postulates 1 and 2. There is complete rent dissipation, and the rent dissipation is socially wasteful (the consumer price remains to c_0 after the innovation). The model is extreme in several respects. First, product differentiation, say, could lead the preempted firm to follow suit in the long run. Second, the preempted firm does not exit in spite of zero gross profit, an assumption inconsistent with the existence of fixed costs. Third, that the outcome is completely wasteful relies on our assumption that the innovation is not drastic. All these features can easily be incorporated into the model to enhance its realism.

B. *Delayed Joint Adoption*

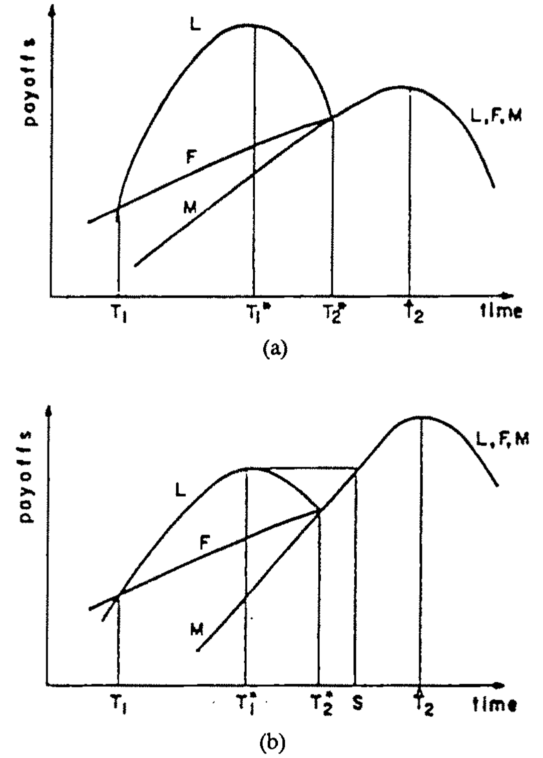
To give a simple example of delayed adoption, consider the following discrete time model: Each duopolist initially makes profit $\Pi > 1$ per period. The current cost of adopting a new technology, $\tilde{p}$, is constant over time (this violates our previous assumptions, but in an irrelevant way,) with $1 < \tilde{p} < (1+r)/r$ where r is the per period rate of interest. If only one firm (the "leader") has adopted the technology, its flow rent is $\Pi + 1$ and its rival (the "follower") has flow rent $\Pi - 1$. If both firms have adopted, their flow rent is Π again. Thus, the innovation serves merely to transfer profits from one firm to the other. At each date, each firm chooses whether to adopt (if it has not adopted yet) depending on the previous history.

There are several perfect equilibria in this game. We focus on the Pareto-inferior and Pareto-superior ones. Note first that one always has immediate reaction: If one firm adopts at t, then the other adopts at $t+1$, because the flow profit associated with adoption, equal to 1, exceed the interest $\tilde{p}r/(1+r)$ on the adoption cost. *The Pareto-inferior* (*preemptive*) *equilibrium* has each firm adopt at each date (conditional on not having adopted before) independently of whether its rival has yet adopted. In this equilibrium, both firms adopt at date 0 and have payoff $\Pi\cdot((1+r)/r)-\tilde{p}<\Pi\cdot((1+r)/r)$. Firms are made worse off by the introduction of the innovation ("super" rent dissipation). The *Pareto-superior* (*delayed adoption*) *equilibrium* has each firm adopt only if its rival has adopted before. Thus, adoption never takes place. Each firm's payoff is $\Pi\cdot((1+r)/r)$. This is an equilibrium because $\tilde{p}>1$. There is no (super) rent dissipation.

It is interesting to note that in the open-loop information structure (firms do not observe their opponent's choice), the unique equilibrium outcome is the preemptive one described above: given that the rival's adoption date cannot be influenced, it is a dominant strategy to adopt as early as possible, because the extra flow profit exceeds the interest saved by delaying adoption by one period.

This stylized example may be a good model of an arms race, but it is not a good representation of market competition. However, the example does contain all the intuition required to understand economically more interesting examples. Adoption in economic contexts is studied in our 1985 paper (see also Michael Katz and Carl Shapiro, 1985). Let $L(t)$ and $F(t)$ denote the equilibrium payoffs of both the leader and the follower as a function of the time of *first* adoption. These curves are displayed for continuous time in Figure 1. Let $T_2{}^*$ denote the follower's optimal reaction time given that the leader has adopted at $t\leq T_2{}^*$ (in the absence of goodwill effects or the like, $T_2{}^*$ is independent of t). For $t\geq t_2^*$, the follower follows suit immediately so that $L(t)=F(t)$. The payoff to simultaneous adoption is denoted by $M(t)$.

Figure 1a is similar to imitation deferring innovation. Only the preemption equilibrium, with dates T_1 for the leader and $T_2{}^*$ for the follower, exists. Figure 1b is similar to delayed joint adoption. In addition to Pareto-inferior preemption equilibrium at $(T_1,T_2{}^*)$, there also exists a Pareto-superior, delayed joint adoption equilibrium at $\hat{T}_2$.



(a)

(b)

FIGURE 1

Is this joint adoption equilibrium plausible? F. M. Scherer (1980) argues that something of this sort may explain the delay of the American automobile industry in introducing seatbelts. Our model predicts that the joint-adoption outcome is more plausible in an industry where adoption by one firm triggers a fast response by its rival, which may be the case for seatbelts.

Joint adoption also requires that players can and will respond to early adoption by adopting early themselves. This sort of response is ruled out by the open-loop information structure considered by Reinganum. Her results show that in any pure-strategy open-loop equilibrium, one firm adopts at $T_2{}^*$, and the other at the time $T_1{}^*, T_1<T_1{}^*<T_2{}^*$, which is the best response for a firm whose opponent is certain to wait until $T_2{}^*$. This outcome is not an equilibrium when firms can respond to their opponent's actions: since $F(t_1^*)<L(T_1{}^*)$, the firm which is meant to adopt at $T_2{}^*$ would prefer to deviate and adopt just before $T_1{}^*$, thus pre-

empting its opponent. Payoffs in the open-loop equilibrium are intermediate between those in the preemptive and joint-adoption equilibria discussed above.

Of course, both negligible information lags and infinite ones are extreme cases and not exactly descriptive of any market, so one might wonder how game theory has contributed to our understanding of the adoption problem. Our answer is that without the discipline imposed by the requirement that strategy spaces are clearly specified, the crucial role that information lags play in this problem might be overlooked. Also, our initial conjecture was that with negligible information lags, the preemptive outcome would necessarily occur. The exercise of characterizing the set of equilibria led us to discover the possibility of delayed joint adoption, which seems to be not an artifact of the model, but the appropriate prediction in some circumstances.

III. Patent Races

The final situation we consider is that of patent races, where we will again see the key role played by the extent of information lags. The model here is much like that of Section II, Part A, except that "winning" is a multiple-step process. The "race" is a very stylized notion of the *R&D* process: Obtaining the patent requires completing a prespecified number of steps in a fixed order and the first one to do so wins the patent. For simplicity, we further assume that this process is completely deterministic, and takes the following form: spending zero maintains a firm's position, spending 1 increases the position by 1, and spending 3 increases the position by 2. There is no discounting, so that the efficient program is to progress one step per period. Firms move simultaneously, and at the start of each period firms are perfectly informed about their opponents' past progress. This is the model of Fudenberg et al. (1983); Chris Harris and John Vickers (1985) consider a sequential-move version with a more general *R&D* technology.

Dasgupta and Stiglitz conjectured that if one firm had a slightly higher payoff for winning the patent, or a headstart in the race, then in equilibrium it would not only win, but do so while spending no more on *R&D* than if it were a monopolist, that is, there would be no rent dissipation at all. This prediction is radically different than the equilibrium outcome in the model of Section II. The basis for the no-dissipation conjecture is that the favored firm ("*FF*") can break even at an investment speed that would be unprofitable for any opponent. Thus the *FF*'s opponents predict that it will always choose to spend just enough to remain ahead, and so the *FF* is able to proceed unopposed at the low, efficient speed.

The key to this argument is that if another firm did compete for the patent, the *FF* will learn of its activities quickly enough to maintain a lead. This is where the information lags resurface. The no-dissipation argument presupposes that *FF* cannot be "caught napping" by an opponent who suddenly appears on the scene with a commanding lead. In our special model, this requires that *FF* to have a lead of at least 2. More generally, the *FF*'s lead must exceed the most progress that can be made in a single period without spending more than the monopoly value of the resulting position. Thus, if the time between periods (the observation lag) is long, the *FF* must have a substantial initial advantage to capture the monopoly rents, as the observation lag shrinks, the required asymmetry lessens. Information lags favor rent dissipation here, contrary to the model of Section II. Once again, we see that the extent of rent dissipation is highly dependent on the game-theoretic aspects of the situation.

REFERENCES

Baumol, William, Panzar, John and Willig, Robert, *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

Dasgupta, Partha and Stiglitz, Joseph, "Uncertainty, Industrial Structure, and the Speed of *R&D*," *Bell Journal of Economics*, Spring 1980, *11*, 1–28.

Eaton, B. Curtis and Lipsey, Richard G., "Exit

Barriers are Entry Barriers: The Durability of Capital as a Barrier to Entry," *Bell Journal of Economics*, Autumn 1981, *12*, 593–604.

Fisher, Franklin, "The Social Costs of Monopoly and Regulation: Posner Reconsidered," *Journal of Political Economy*, April 1985, *93*, 410–16.

Fudenberg, Drew et al., "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, June 1983, *22*, 3–31.

Fudenberg, Drew and Tirole, Jean, "Preemption and Rent Equalization in the Adoption of a New Technology," *Review of Economic Studies*, June 1985, *52*, 383–401.

______ **and** ______, "A Theory of Exit in Duopoly," *Econometrica*, July 1986, *54*, 943–960.

Harris, Chris and Vickers, John, "Perfect Equilibrium in a Model of a Race," *Review of Economic Studies*, April 1985, *52*, 193–209.

Katz, Michael and Shapiro, Carl, "Perfect Equilibrium in a Development Game with Licensing or Imitation," mimeo., Princeton University, 1985.

Maskin, Eric and Tirole, Jean, "A Theory of Dynamic Oligopoly, I: Overview and Quantity Competition with Large Fixed Costs," Department of Economics Discussion Paper No. 1270, Harvard University, 1986.

Posner, Richard A., "The Social costs of Monopoly and Regulation," *Journal of Political Economy*, August 1975, *83*, 807–27.

Reinganum, Jennifer, "On the Diffusion of New Technology: A Game-Theoretic Approach," *Review of Economic Studies*, July 1981, *48*, 395–406.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2nd ed., Chicago: Rand-McNally, 1980.

Informational Asymmetries, Strategic Behavior, and Industrial Organization

By PAUL MILGROM AND JOHN ROBERTS*

One of the most active and exciting areas of economic research over the last several years has been the use of noncooperative games of incomplete information to model industrial competition. This work has yielded not only a remarkable number of papers but also several new insights on and explanations of fundamentally important issues. The purpose of this paper is to attempt an appreciation and evaluation of this work. Because most of our individual and joint work since about 1979 has been in this mode, it will come as no surprise that we are proponents of this line of research. However, there are several questions and potential problems that we see as arising in connection with this methodology, and we will attempt to address these.

First, a disclaimer. We are not attempting a survey of the applications of asymmetric information games (AIG) to industrial organization, although we will refer in a highly selective fashion to a number of prominent strands in this literature. (In particular, where any references are provided at all, they are typically only to the earliest contributions to a subject.) Even more, we do not deal with work in which informational asymmetries are important but the analysis is not game theoretic (for example, search and price dispersion, or the early work on the lemons problem and on moral hazard and adverse selection in insurance markets) or with game-theoretic treatments that assume complete information.

I. AIG Methods and Applications

To get an idea of the role of informational asymmetries in strategic behavior, consider three simple card games. In the first, each player is dealt five cards face up, the players make any bets they want, and then the best hand wins. In the second, each player receives five cards, some of which are dealt face up and the rest face down. Without looking at their hole cards, the players make their bets, then the cards are turned face up and the best hand wins. Finally, the third game is like the second except that players can look at their hole cards. Again there is betting, the hidden cards are revealed, and the best hand wins.

The first game is one of complete (and perfect) information. Everyone knows everything, and as long as we assume that people prefer more money to less, it is fairly trivial to figure out what will happen: there will certainly be no betting, and probably no one will bother to play! Clearly, not all games of complete information are either so uninteresting (witness chess) or so lacking in explanatory power—especially if we consider nonzero sum games and, even more, games with an explicit dynamic structure (Drew Fudenberg and Jean Tirole, 1986a). However, in its informational structure, this game typifies both the sort of game theory that is discussed in intermediate micro texts and, indeed, most of standard microeconomic theory itself.

The second game has uncertainty/informational incompleteness, but no informational asymmetries. Its informational structure puts it in the domain of decision theory and the economics of uncertainty. Games of this sort are useful models for studying such issues as insurance, risky investments, and learning (especially if we revise the game to have the hole cards revealed one at a time,

*School of Organization and Management and Department of Economics, Yale University, New Haven, CT 06520, visiting the University of California-Berkeley; and Graduate School of Business, Stanford University, Stanford, CA 94305-5015, respectively. We gratefully acknowledge the comments of David Kreps and Robert Wilson, and the financial support of the John Simon Guggenheim Memorial Foundation and the National Science Foundation.

with betting after each is shown). However, its play would not generate any interesting forms of strategic behavior.

The third game involves informational asymmetries: while there is some publicly available information, each player is privately informed about his or her hole cards. (In fact, the informational structure of this game, in which the probability distribution over what the particular private information of the various players could be is common knowledge, corresponds very closely to that in the asymmetric information game models used in most applications to industrial organization.) The existence of this private information can obviously lead to interesting strategic play: bluffing, signaling, reputation building, etc. It is also the reason why poker is of enduring popularity.

As this example is meant to suggest, recognition of informational asymmetries and the strategic possibilities they engender can yield models that begin to capture the richness of behavior that marks the real world. This is the great advantage of these methods: they permit us to model, and thereby start to understand, phenomena that made no sense in terms of complete information analyses or ones based on incomplete but symmetric information (uncertainty).

Perhaps the clearest example of this is predatory pricing. In 1980, the only fully consistent analyses of predatory pricing in the literature (for example, John McGee, 1958, 1980) indicated that predatory pricing could not be expected to succeed, that it was thus not part of a rational competitive strategy, that apparent instances of predatory pricing were consequently likely to be either mistakes or misinterpretations, and that legal prohibitions on predation serve chiefly to protect inefficient firms from the desirable effects of competition. Although these conclusions ran counter to much of the conventional wisdom in the field of industrial organization and left a disconcerting number of mistakes and misperceptions to explain, the logic leading to them seemed compelling. And because no mere fact ever was a match in economics for a consistent theory, these ideas began to represent the basis of a new consensus.

However, these analyses rested on an implicit assumption of symmetric information, and this assumption is crucial. Recent studies by a number of authors (see Roberts, 1987, for a survey and complete references) have relaxed this assumption in various ways and reversed the earlier findings. Pricing below the short-run optimal level aimed at deterring entry, inducing exit, or disciplining rivals so that they compete less aggressively can be part of a rational strategy in the presence of realistic informational asymmetries. Thus we should expect to see predatory behavior being adopted if it is not effectively deterred by legal prohibitions. Moreover, the mechanisms by which these effects come about in these theories correspond well to suggestions found in earlier, less formal discussions of predation in the industrial economics and legal literatures. At the same time, these new analyses indicate that the legal tests that have been proposed for establishing whether predation has occurred may be completely inappropriate in that they may fail in either direction.

Most of the recent, asymmetric information game models of predatory pricing involve generalized signaling: there is an underlying parameter that is of interest to one player (the "receiver") but is not directly observed by this player, and the (costly) actions of the other player (the "signaler" or "sender") can affect the observations made by the receiver. By the choice of actions the sender can thus influence the inferences the receiver makes about the parameter's value and, correspondingly, the receiver's choice of actions. This formulation subsumes the original Spence-type signaling model, in which the signaler (there, the worker) knows the value of the parameter (his or her productivity) and so can condition his or her choice (of education level) on this information. In these circumstances, observation of the sender's choice may allow the receiver to infer the sender's information. It also encompasses "signal jamming" models, in which neither player is informed about the value of the parameter. In these models, the actions of the signaler are not observed by the receiver directly; instead they affect the distribution of a variable that is observed by the receiver and from which he or she must try to infer the value of the parameter. As well, dynamic formulations in which the signaler acts repeatedly and information is revealed over time (perhaps to a sequence of receivers) are also included.

In signaling models of predatory pricing, the sender is the predator firm, the receiver is either the current rival on which the predation is practised or potential rivals that can observe the predator's current behavior, the parameter is a variable that influences the receiver's profit from continued operations, and the signal is the predator's price. The models of Roberts (1986) and Garth Saloner (1986) of predation against a single opponent are Spence-type signaling ones in which the predator is privately informed about demand or cost. The signaling firm is led to lower its price in an attempt to suggest that the value of the parameter is such that either continued competition by the receiver will be unprofitable—thereby inducing exit (Roberts) or encouraging acceptance of a merger offer (Saloner)—or that a reduced level of output would be optimal for the receiver. (See also David Scharfstein, 1984.) Fudenberg and Tirole (1986b) consider a situation where the two firms are symmetrically informed about the random demand, but the predator has an incentive to increase output or lower price unobservably and thereby attempt to bias the receiver's estimates of profitability. Finally, David Kreps and Robert Wilson (1982), our paper (1982b), and Fudenberg and Kreps (1986) consider dynamic models in which the privately informed sender adopts predatory behavior against early entrants, even though it is directly unprofitable, in order to build a reputation for aggressive responses to entry that deters future challenges. See also David Easley, Robert Masson, and Robert Reynolds (1985), where the reputation being built is for having markets into which entry is unprofitable because demand is weak.

Equilibrium in signaling models involves the receiver's having correct conjectures regarding the sender's actions (as a function of the sender's information) and accounting for these in making inferences about the parameter. Thus, both in models in which the sender is uninformed about the parameter and in ones where the sender knows its value, the receiver's estimates are not systematically biased by the signaling behavior. In fact, to the extent that the equilibria are separating (involve a one-to-one map between the parameter's value and the receiver's observations), the receiver will correctly infer the value of the parameter and will thus, in effect, be acting as if he had access to the sender's information. Yet the sender's actions are typically distorted from their full-information levels because the receiver interprets his observations in light of the sender's incentives to attempt to influence his inferences. Thus, for example, in the Roberts model, if the signaler produced the full-information output rather than the equilibrium one, the receiver would interpret the resulting observation as indicating that demand is stronger than it actually is. As a result, the receiver would be less likely to exit, and if it stayed in, it would produce at the higher level corresponding to strong demand.

This property of separating equilibrium has welfare implications: not only is the price lowered during the predatory episode, but there need be no more exit or restriction of output by the prey than there would have been if predation had been effectively forbidden. However, this does not imply that the predatory behavior is socially desirable, because the threat of predation (even if it will fail to induce exit) will deter entry. This is very clear in the Kreps-Wilson and Milgrom-Roberts dynamic reputation models, but is also true in the static models. Thus, legal concern with predation may still be warranted. At the same time, these models indicate that predation may occur without the predator ever pricing below marginal cost or in other ways meeting the tests for predatory behavior often advocated in the literature.

The idea of signaling has also proven fruitful in studying a number of other situations involving pricing under imperfect competition. An early example was our (1982a) rationalization of limit pricing, in which low-cost incumbents are led to price below the short-run monopoly level to signal credibly that the entrant will find their markets unprofitable targets for entry (see our 1982a paper; again see Roberts, 1987, for further references). George Mailath (1985) has used signaling to explain the phenomenon of price wars occurring early in an industry's history: each firm expands output in an attempt to make its competitors believe that its cost are low and that it thus should have a large market share. In a closely related model, Michael Riordan (1985) has used signal jamming to rationalize conjectural variations. A further example is Kyle Bagwell's (1986) signaling explanation of introductory sales: the firm seeks to signal that its costs are low and thus that it will be worthwhile for customers to shop at this firm in the future because its low costs will lead it to set low prices then as well. Our recent treatment (1986a) of both pricing and image advertising as signals for unobservable product quality carries these methods into a multidimensional context.

Models that explicitly incorporate private information but are not of the signaling variety have also been widely used. For example, they have proven useful in examining the incentives to reveal cost, demand and other information to suppliers (Milgrom and Robert Weber, 1982a, b) and to competitors through trade associations and the like (William Novshek and Hugo Sonnenschein, 1982). They have been applied to the problem of maintaining cartel agreements when there are problems in detecting cheating (Edward Green and Robert Porter, 1984, and Dilip Abreu et al., 1986). The burgeoning literatures on nonlinear pricing, priority pricing, durable goods monopoly, auctions, bargaining, etc., are all based in this methodology. Finally, theories of contracting in the presence of informational asymmetries, which are increasingly being based in the methodology of incomplete information game theory, are contributing mightily to our understanding of issues of procurement, regulation, vertical integration, and the functioning and design of economic organizations and institutions more generally. Moreover, all this has been accomplished essentially in the last five years.

The methods of incomplete information game theory have thus allowed us to model formally, often for the first time, issues that are central to industrial organization. Moreover, the behavior that emerges in equilibrium from these models begins to capture, again often for the first time, something of the richness of observed behavior. Indeed, this work holds some promise of yielding a partial and much-belated realization of some of the hopes that were expressed in the early years of game theory that its use would lead to a resolution of the problems of oligopoly and imperfect competition. This certainly (to us) justifies them and their use in industrial organization research.

II. Questions about the AIG Methodology

The questions that can and do arise about these methods seem to us to fall under five headings: the assumption that equilibrium behavior will prevail and the related but more fundamental rationality assumptions; the assumed common knowledge base; robustness of the results; multiplicity of equilibria; and empirical implementation and testing. We discuss each of these in turn.

Almost all of economic theory is equilibrium analysis, and this work is no exception: predictions arise only once equilibrium behavior is assumed. Yet the assumption of equilibrium seems more demanding in many of these incomplete information games than it does in, say, competitive partial-equilibrium models. There seem to be two possible reasons for this.

The first of these relates to the relative specificity of the two types of models. Most standard models of price determination are remarkably incomplete: the timing of actions, the information available to agents when they act, and the consequences of not adopting equilibrium behavior are rarely modeled, and many treatments even leave out such seeming fundamentals as who selects prices, and how the supplies and demands expressed to the market get transformed into actual transactions. This incompleteness facilitates accepting the equilibrium assumption because the model gives nothing else on which to focus. At the same time, it encourages us to comfort ourselves with vague and often unarticulated stories about processes of adjustment that somehow lead to equilibrium. In contrast, the methodology of extensive games forces AIG models to be much more specific and complete. First, this calls to our attention the possibilities of not adopting equilibrium behavior. Secondly, it invalidates appeals to learning through repetition or to adjustment of play over time toward equilibrium, because if there is repeated play, this should be modeled and the resultant game analyzed on its own. As is well known, the equilibria of the repeated-play game may differ substantially from those of the one-shot version.

Of course, to the extent that the assumption of equilibrium seems easier in more standard models because they assume away complicating factors that actually may be important, a finding that the equilibrium assumption is relatively more problematic in asymmetric information-strategic models is hardly a criticism of these models.

The second reason relates to the complexity of the coordination problem in these models. There are often multiple equilibria in AIG models, and one player's adopting his or her strategy from one equilibrium and the others' playing their strategies from a second equilibrium typically does not constitute equilibrium behavior. This problem, of course, is not unique to AIG models: consider the Battle of the Sexes. Still, even if everyone can figure out what strategy n tuples are equilibria, unless there is a unique equilibrium, there are real problems in ensuring that everyone focuses on the same one. There are various stories that game theorists tell in such situations (see Kreps, 1986, for an exposition of these), but none are fully satisfactory. Of course, all these problems disappear if there is a single equilibrium. (We return to the multiplicity issue below.)

Even if the coordination problem can be overcome and if one grants that rational actors would adopt equilibrium behavior, one might still more fundamentally question the rationality assumption as it appears in AIG models.

There is no denying that the sort of inferences, calculations, and forecasts that agents are making in the equilibria of AIG models involve much more sophistication than, say, the agents in an Arrow-Debreu world of complete, competitive markets must show when the equilibrium prices are given. Of course, this latter standard is the extreme, and the demands on rationality increase as soon as we move from this world to ones without complete, perfect markets and symmetric information. Still, equilibrium in AIG models does seem to involve another quantum leap beyond, say, rational expectations models with informative prices. As game models, AIG treatments require players to act as if they anticipate fully the often complex responses of the other players. Further, in AIG models these responses depend on subtle inferences that competitors draw, often by very intricate reasoning, from their conjectures about others' behavior and their observations. The strategic players' decision problems therefore are much more difficult than in more standard models. Correspondingly, the assumption of agents' finding and adopting equilibrium behavior becomes that much more implausible.

In this regard, it seems to us that an appropriate test of the assumptions of rationality and equilibrium is the standard one: if their use aids our understanding, leads to accurate descriptions, facilitates prediction, and generates useful recommendations, then use the assumptions until something better comes along. Nevertheless, the descriptive accuracy of the super-rationality assumption does seem minimal. Indeed, if it did describe actual individuals, the outcome of the game of chess would be totally determinate and obvious, and its play would be as exciting as Tic-Tac-Toe. Given that AIG models have not been around long enough to have yet stood the above test on many occasions, this lack of descriptive accuracy is troubling. Moreover, there do seem to be important phenomena, especially in the economics of organization, that are very hard to explain without retreating at least from the assumption that transferring information, assimilating it, calculating, and deciding can be done instantaneously and without cost.

There have, of course, been some recent models in the AIG framework that include boundedly rational agents. For example, our 1982b model of predation and the Kreps et al. (1982) treatment of the finitely repeated Prisoner's Dilemma involve the possibility of agents' using simple rules of thumb to guide their behavior, and more recent work by a number of game theorists has investigated situations where strategies must be implementable by finite automata or involve limited memory. However, hyperrational agents still play a crucial role in all this work. In the work with which we have been associated, rational agents in equilibrium are led by very complex reasoning to mimic the rule-of-thumb players, while in the later work the machines used to implement strategies are selected by rational agents making the usual sorts of forecasts, inferences, and calculations.

The problem, of course, is that we as yet have little agreement on how to model more descriptively accurate forms of rational behavior, little faith that we can find hypotheses on behavior that will be as tractable and powerful as maximization and equilibrium, and a general fear that by renouncing our standard methods we will forfeit elegance and, in return, get only *ad hockery.*

A related complaint about AIG models is the common knowledge assumption. In applications to industrial organization, it is typically assumed that the private information is of small dimension (for example, cost function parameters) and that the distribution of possible values for this information, as well as everything else in the model, is common knowledge. (Intuitively, an event is common knowledge if each player knows it has occurred, each knows that each knows this, each knows that each knows that each knows this, *ad infinitum.*) The objection to this that one hears is that this is no more realistic than assuming that actual values are common knowledge, and, by implication, analyses based on such an assumption are at least as suspect as ones assuming complete information.

It seems that there are two possible interpretations of this complaint. The first is that assuming that beliefs over the values of the underlying parameter are common knowledge is too simplistic, because there might reasonably by uncertainty about beliefs. Formally, of course, there is no need to assume that the type space is so simple: your type could involve not only your costs but also, for example, your beliefs about what others believe your costs are. Such relatively complicated spaces have in fact been successfully used in applications to industrial organization (see our 1982b paper, appendix B); however, generally they will be incompatible with obtaining the sort of monotonicity properties that have proven so fruitful in signaling models.

The second interpretation relates to the concerns with rationality. The type spaces in AIG models rapidly become extremely complicated mathematical structures as the level at which the uncertainty is assumed to lie is pushed back. For example, if there are two matrices that might give the payoffs, then the possible beliefs over which of these prevails correspond to points in the unit interval, beliefs over beliefs are increasing functions from [0,1] into itself, the next level of beliefs are the measures on this space of functions, and so on. Moreover, whatever the assumed type space, to use AIG methods the distribution of types must be taken to be common knowledge, as must, in fact, the full game tree and the payoffs at each terminal node, the number of which is a function of the number of types. To assume that real people make calculations over such complex spaces seems again to strain the limits of credibility.

This latter point might be thought to be a problem that is particular to the AIG methodology. However, this is not quite the case. One can argue that *any* game-theoretic method of analyzing rational behavior in multiperson decision problems must start from an assumption of what is common knowledge among the agents (see, for example, Wilson, 1986). Thus, AIG models are in no way special in this regard; instead, this methodology, in which the common knowledge assumptions are completely explicit, has made the more general necessity of such assumptions apparent.

While we agree with this position, some of our recent work suggests that it may be important to study the extent to which this point is valid once one moves to an assumption of bounded rationality. In particular, in our 1986b paper, we consider a decision maker whose rationality is decidedly bounded; in particular, he or she does not even know the type space, let alone have a common knowledge distribution over it reflecting beliefs. Yet this decision make can do very well in certain situations by being skeptical and by inducing competition between informed interest parties.

Another potentially disturbing aspect of these models that is less obvious is the apparent sensitivity of the results to alterations in what one might think is fine structure of the models. This is especially clear in finitely repeated games. Kreps-Wilson and our paper (1982b) have shown that the introduction of a tiny bit of private information into such a model can radically change its equilibrium outcomes as the finite horizon becomes "long." However, this technique may be too powerful: Fudenberg and Eric Maskin (1986) have shown that by introducing the right kind of informational asymmetries, one can obtain a Folk Theorem result that almost anything can be made an equilibrium (see also John Ledyard, 1986). What is needed is some way to determine which informational asymmetries have survival value in the sense that if people ascribe positive probability to a variety of possible types, then those with survival value are the ones whose behavior is mimicked (as are the tit-for-tat players in Kreps et al.). See our paper (1982b) and Robert Aumann and Sylvain Sorin (1986) for beginnings in this direction.

Even in one-shot games there are important discontinuities. For example, Mailath has shown that there is a unique separating sequential equilibrium in our limit-pricing model with a continuum of possible values for the incumbent's cost. It involves all but the highest-cost type of incumbent producing more than the monopoly output. This equilibrium is unaffected by changes in the relative likelihood of the incumbent's possible types, so long as the support of the beliefs is unchanged. But suppose that the weight becomes concentrated in the limit on the lowest cost type. At this limit, sequential equilibrium requires a discrete jump in the firm's output choice to the simple monopoly price and output. This same example illustrates the very important point that the addition and deletion of initial nodes (types) having zero prior probability can radically affect the solution of these models. Similarly, the inclusion of the option to take actions that might, on first blush, appear to be dominated—such as publicly burning money—can affect the solution (see our 1986a paper).

The question here is whether these features are artifacts of the models or whether they correspond to something real. If it is the latter (as we suspect it is—the range of possibilities that people consider will affect their decisions and actions), then this sensitivity of behavior is unfortunate for those who hope to draw easy general conclusions, but must be faced. In particular, since one doubts that everyone is always certain that everyone else is super-rational and, more generally, that the model of the world they are using is absolutely accurate, it is crucial to investigate models including informational asymmetries. In this regard, the study of how reasonable forms of bounded rationality affect the sensitivity of models assuming hyper-rationality seems especially important.

A further complaint against the AIG models that have been used in industrial organization is that they admit such a great multiplicity of equilibria that, while they might possibly explain certain phenomena, they are of limited value for prediction because so many patterns of behavior are consistent with equilibria. This is a criticism of which both formal game theorists and those interested in applying AIG models have been very aware, and members of both groups have devoted significant effort to meeting it.

The feature that makes the multiplicity problem especially acute in AIG models compared to complete information games is the multitude of beliefs about the relative likelihood of various events that can be held *off the equilibrium path* and can be consistent with equilibrium. Sequential equilibrium requires that, at each decision point (and not just those reached under equilibrium play), each player find that continuing to use his or her equilibrium strategy is optimal given his or her beliefs about what has happened so far and what the others know. In situations without informational asymmetries, this subgame perfection requirement often has the power to delimit actions quite narrowly. However, in AIG models, it is typically the case that the set of optimal actions varies widely as we alter the players' beliefs about how play has proceeded and what information the others may have. Further, Bayesian updating gives no restriction on beliefs off the equilibrium path, since in such situations we would be conditioning on probability zero events. Thus, a wide variety of beliefs —and a correspondingly wide variety of optimal actions—are consistent with equilibrium.

Of course, if the indeterminacy were confined to situations that are not observed in equilibrium (and if one decides to focus on equilibria), there would be no great problem. However, this cannot be the case. One of the great virtues of game-theoretic methodology is the requirement that behavior be specified in *all* eventualities, not just under the particular circumstances corresponding to some putative equilibrium. Then equilibrium is determined endogenously by considering the implications of deviating from the specified behavior. These implications are fully determinate exactly because players' strategies specify what actions to take at *every* decision point. Thus, in effect, off-the-equilibrium-path behavior determines equilibrium behavior. Correspondingly, in AIG models there can be many equilibria, each supported by different beliefs off the equilibrium path and by the behavior these generate.

Given this diagnosis, most of the effort aimed at overcoming the multiplicity problem has been directed toward narrowing the set of out-of-equilibrium beliefs that are to be considered to represent *reasonable* inferences to have made after observing others' choices (see, for example, Kreps). In some circumstances, relatively simple and intuitively appealing conditions suffice to reduce the set of equilibria significantly, and even to generate uniqueness. For example, in signaling games it is often enough to assume that if a particular message could never be part of a best response for some type of sender but could for others, then if this action is observed, receivers do not attribute it to the sender type for whom it is never a best response. However, in other situations, either the required conditions are very hard to understand or simply are not yet known (for example, bargaining with private information on both sides). Moreover, to the extent that obtaining a small set of equilibria involves agents' making particular, highly sophisticated, extremely subtle inferences from observations, the questions raised about the equilibrium and rationality assumptions arise again.

Finally, there is the issue of testing these models. There has, to date, been relatively little empirical work based on AIG models. In part, this may simply be a matter of time. These models are new and have for the most part been developed by theorists; they may simply not yet have reached the empiricists' agenda. However, it seems that there are inherent difficulties in testing. Some of these arise from the multiplicity of equilibria, but the more central one is that the central object in the theory is, by its very nature, unobservable. How, for example, does one obtain information on what a firm's beliefs about rivals' costs were when it took a particular pricing or entry decision? Note that the sensitivity problem discussed above exacerbates this difficulty, because the predicted results can depend so finely on both the distributions over private information and the fine details of the modeling.

This suggests two approaches: careful case studies and experimental work. The former is clearly very costly, which may be why, to our knowledge, relatively few such studies have been attempted. However, those that have been done (for example, Mark Wolfson, 1985, on contracting; Michael Staten and John Umbeck, 1982, on shirking in labor markets; Robert Porter, 1983, on cartel maintenance) are generally supportive of the theory, as are various other studies not specifically based on theories involving informational asymmetries (for example, Malcolm Burns, 1986). Meanwhile, a significant part of the experimental work on bargaining, auctions, and various market institutions incorporates private information, and while little work directly aimed at examining the influence of information asymmetries on strategic behavior in industrial organization settings has yet been done, those studies that do exist (for example, Ross Miller and Charles Plott, 1985; Colin Camerer and Keith Weigelt 1986) again tend to give some support to the theory. These results, and the exciting insights that the theory has offered, justify the attention that AIG models in industrial organization theory have received and, we trust, will continue to receive.

REFERENCES

Abreu, Dilip, Pearce David and Stacchetti, Ennio, "Optimal Cartel Equilibria with Imperfect Monitoring," *Journal of Economic Theory*, June 1986, *39*, 251–69.

Aumann, Robert and Sorin, Sylvain, "Bounded Rationality and Cooperation," paper presented at annual meeting of the American Economic Association, New Orleans, December 29, 1986.

Bagwell, Kyle, "Introductory Price as a Signal of Cost in a Model of Repeat Business," Studies in Industry Economics, No. 130, Department of Economics, Stanford University, 1985.

Burns, Malcolm, "Predatory Pricing and the Acquisition Cost of Competitors," *Journal of Political Economy*, April 1986, *94*, 266–96.

Camerer, Colin and Weigelt, Keith, "Experimental Tests of a Sequential Equilibrium Reputation Model," mimeo., Graduate School of Business, New York University, September 1986.

Easley, David, Masson, Robert and Reynolds, Robert, "Preying for Time," *Journal of Industrial Economics*, June 1985, *33*, 445–60.

Fudenberg, Drew and Kreps, David, "Reputation and Multiple Opponents I: Identical Entrants," mimeo., Graduate School of Business, Stanford University, May 1986.

______ **and Maskin, Eric,** "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, May 1986, *54*, 533–54.

______ **and Tirole, Jean,** (1986a) *Dynamic Models of Oligopoly*, New York: Harwood Academic, 1986.

______ **and** ______, (1986b) "A 'Signal-Jamming' Theory of Predation," *Rand Journal of Economics*, Autumn 1986, *17*, 366–78.

Green, Edward and Porter, Robert, "Noncooperative Collusion under Imperfect Information," *Econometrica*, January 1984, *52*, 87–100.

Kreps, David, "Out of Equilibrium Beliefs and Out of Equilibrium Behavior," Working Paper, Graduate School of Business, Stanford University, September 1986.

______ **and Wilson, Robert,** "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, *27*, 253–79.

______ **et al.,** "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, August 1982, *27*, 245–52.

Ledyard, John, "The Scope of the Hypothesis of Bayesian Equilibrium," *Journal of Economic Theory*, June 1986, *39*, 59–82.

McGee, John, "Predatory Price Cutting: The Standard Oil (N.J.) Case," *Journal of Law and Economics*, October, 1958, *1*, 137–69.

______, "Predatory Pricing Revisited," *Journal of Law and Economics*, October 1980, *23*, 289–330.

Mailath, George, "Separating Equilibria in Signaling Games: Incentive Compatibility, Existence with Simultaneous Signaling, Welfare and Convergence," unpublished doctoral dissertation, Princeton University, 1985.

Milgrom, Paul and Roberts, John, (1982a) "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis," *Econometrica*, March 1982, *50*, 443–59.

______ **and** ______, (1982b) "Predation, Reputation and Entry Deterrence," *Journal of Economic Theory*, August 1982, *27*, 280–312.

______ **and** ______, (1986a) "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, August 1986, *94*, 796–821.

______ **and** ______, (1986b) "Relying on the Information of Interested Parties," *Rand Journal of Economics*, Spring 1986, *17*, 18–32.

______ **and Weber, Robert,** (1982a) "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, *50*, 1089–122.

______ **and** ______, (1982b) "The Value of

Information in a Sealed-Bid Auction," *Journal of Mathematical Economics*, June 1982, *10*, 105–14.

Miller, Ross and Plott, Charles, "Product Quality Signaling in Experimental Markets," *Econometrica*, *53*, July 1985, 837–72.

Novshek, William and Sonnenschein, Hugo, "Fulfilled Expectations Cournot Dupoly with Information Acquisition and Release," *Bell Journal of Economics*, Spring 1982, *13*, 214–18.

Porter, Robert, "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886," *Bell Journal of Economics*, Autumn 1983, *14*, 301–14.

Riordan, Michael, "Imperfect Information and Dynamic Conjectural Variations," *Rand Journal of Economics*, Spring 1985, *16*, 41–50.

Roberts, John, "Battles for Market Share: Incomplete Information, Aggressive Strategic Pricing, and Competitive Dynamics," in T. Bewley, ed., *Advances in Economic Theory*, Cambridge: Cambridge University Press, 1987.

______, "A Signaling Model of Predatory Pricing," *Oxford Economic Papers* November 1986, *38*, Suppl., 75–93.

Saloner, Garth, "Predation, Merger and Incomplete Information," Working Paper No. 383, Department of Economics, Massachusetts Institute of Technology, September 1986.

Scharfstein, David, "A Policy to Prevent Rational Test-Market Predation," *Rand Journal of Economics*, Summer 1984, *15*, 229–43.

Staten, Michael and Umbeck, John, "Information Costs and Incentives to Shirk: Disability Compensation of Air Traffic Controllers," *American Economic Review*, December 1982, *72*, 1023–37.

Wilson, Robert, "Competitive Strategies in Business," Working Paper, Graduate School of Business, Stanford University, September 1986.

Wolfson, Mark, "Empirical Evidence of Incentive Problems and Their Mitigation in Oil and Gas Tax Shelter Programs," in John Pratt and Richard Zeckhauser, eds., *Principals and Agents: The Structure of Business*, Boston: Harvard Business School Press, 1985.

Exchange Rate Management: What Role for Intervention?

By PETER B. KENEN*

Economists are fond of stylized facts—those most striking features of a situation that require explanation, or perhaps those features we can model easily. I will start with some stylized opinions about the costs and benefits of exchange rate management.

In the years right after the shift to floating exchange rates, we heard much talk about monetary autonomy. Floating rates, it was said, would allow each government to control its own money supply and use it to pursue its own policy objectives. Under the Bretton Woods regime, by contrast, there had been two constraints on monetary autonomy.

A normative constraint was imposed by the need to keep inflation rates and interest rates in line with those of the United States, to avoid a persistent deficit or surplus in the balance of payments, and an eventual exchange rate change. A mechanical constraint was imposed by the need to intervene in the foreign-exchange market to keep exchange rates pegged. An official purchase of foreign currency raises the money supply, just like an open market purchase of domestic bonds, and a sale of foreign currency reduces it.

Most central banks tried to sterilize their interventions—to offset the money-supply effects by open market operations. But capital mobility was high, even then, and limited the effectiveness of sterilization; an open market sale of domestic bonds attracted a capital inflow, forcing the central bank to buy more foreign currency. Economists disagreed about the net effect, represented by the size of the offset coefficient, which compared the capital inflow to the open market sale and thus measured the additional intervention induced by any attempt to sterilize previous intervention. We continue to debate that issue but cast in different terms—the degree of substitutability between assets denominated in different currencies or, equivalently, the size and stability of the so-called risk premium.

At that early stage, then, the shift to floating rates was seen to reflect other governments' dissatisfaction with their subservience to U.S. monetary policy, as well as the first stirrings of official interest in monetary targeting. The first wave of discontent with floating rates was also related to monetary policies, but this time to the side effects of having more autonomy. Floating rates were blamed for vicious circles, in which depreciation caused inflation that produced additional depreciation, preventing a change in the nominal exchange rate from having a long-lasting effect on the real rate.

It is now commonly acknowledged that vicious circles cannot spiral on unless they are accommodated by monetary policy, and they may not even start in the absence of inflationary expectations. This brings us up to date. Now that inflationary forces have abated, floating exchange rates are criticized for having excessive effects on real exchange rates, not the deficient effects that were the focus of earlier concern. This was the main complaint about the appreciation of the dollar from 1981 through 1985, which impaired the competitiveness of U.S. industry and contributed to the buildup of protectionist pressures. Those pressures, in turn, were the chief reason for the Plaza Communique of September 1985, in which the G-5 countries sought to talk the dollar down but backed

[†]*Discussants*: Allan H. Meltzer, Carnegie-Mellon University; Maurice Obstfeld, Columbia University.

*Walker Professor of Economics and International Finance and Director of the International Finance Section, Princeton University, Princeton, NJ 08544.

their talk by threating to sell it down by concerted intervention.

Proponents of exchange rate management hailed the Plaza Agreement as the start of a new era in international monetary cooperation, and some have since said that the Plaza Agreement is dead because it disappointed them. Rather than inaugurating close cooperation, it has been followed by disputes about appropriate exchange rates, interest rates, and budget deficits.

The Plaza Agreement marked a sharp change in the attitude of the United States or, more accurately, a change in leadership at the U.S. Treasury. Nevertheless, it should be viewed as an instance of *ad hoc* cooperation undertaken to prevent existing institutions, the GATT system in this case, from crumbling under pressure. Although different in form, it resembled in nature the brief period of close cooperation after the onset of the debt crisis in 1982, inspired by the threat of a banking crisis. This sort of regime-preserving cooperation is gratifying, but a single episode such as the Plaza Agreement cannot be taken to signal a basic change in governments' habits.

The first wave of dissatisfaction with floating exchange rates, when changes in nominal rates were seen to have too little influence on real rates, was not as strong as the second, when changes in nominal rates were seen to be too influential. I venture to suggest a simple reason, in keeping with the spirit of this stylized history. At the time of the first wave of criticism, few officials or economists believed in the feasibility of managing exchange rates, let alone returning to pegged rates. Central banks did not enjoy the credibility they have bought back since. The second wave of criticism has been more effective in mobilizing support for reform of the system because central banks are deemed to have the credibility required to influence market expectations.

Yet there is an odd aspect to this stylized history. Exchange rate management has gained favor recently but intervention has not.

In an early version of his plan for managing the dollar-mark-yen relationship, Ronald McKinnon (1984) wanted to rely primarily on nonsterilized intervention to stabilize exchange rates. In a recent paper, by contrast, he calls for the establishment of target zones but says that central banks should "agree to mutual and symmetrical monetary adjustment to achieve these exchange rate targets" (1986, p. 16). The country with the overvalued currency should reduce the growth rate of its money supply; the one with the undervalued currency should raise its growth rate. McKinnon does not mention intervention. In most presentations of his target-zone proposal, John Williamson assigns a limited role to intervention: Monetary policy should be the principal instrument for keeping rates within their target zones, "reinforced by exchange rate intervention" (1986, p. 166).

What about official views? The Versailles Summit of 1982 created a Working Group to study the effectiveness of intervention. Its conclusions were negotiated carefully and formulated cautiously. It said that intervention has been effective in dealing with disorderly market conditions, by narrowing bid-offer spreads and day-to-day fluctuations. But the Working Group was rather skeptical about the use of intervention for more ambitious purposes—not only the ability of sterilized intervention to alter exchange rates by changing supplies of assets, but even its influence on expectations. From time to time, governments had intervened "when they judged that market participants had not taken full account of fundamental factors...or had lost confidence in the policies of some of the major countries." On occasion, such intervention bought time for market participants to revise their views and showed the authorities' determination to restore confidence. But it was "useless or even counterproductive in the absence of appropriate policy changes" (Working Group, paras. 40, 47–48).

The views of the Working Group were echoed by the Plaza Communique, when it declared that "recent shifts in fundamental economic conditions..., together with policy commitments for the future, have not been reflected fully in exchange markets" and concluded that "in view of the present and prospective changes in fundamentals, some further orderly appreciation of the main non-dollar currencies...is desirable" (paras. 5, 18).

There was much intervention right after the Plaza Agreement, and official accounts gave it more importance than earlier views might have led us to expect. But they gave even more importance to the announcement effects of the Agreement itself and to subsequent policy changes:

> [The] dollar dropped sharply on the day following the G-5 announcement...[It] had already fallen against major foreign currencies by the time the Bundesbank stepped in.... Later the same day, the U.S. authorities conducted their first operation during the period, selling dollars...in a visible manner to resist a rise of the dollar from the lower levels.
>
> During the next few days, there was some skepticism in the market that the lower dollar levels would be maintained.... The Bank of Japan responded with massive dollar sales... [Market] participants came to believe that the authorities were firmly committed to the joint effort....
>
> Late in October the Bank of Japan allowed Japanese money market interest rates to drift higher. It was then that the dollar began to decline particularly sharply against the yen. Many market observers viewed the Japanese actions on interest rates as possibly representing the first of a series of steps to be taken by the G-5 countries....
>
> [*Interim Report*, 1985, pp. 46–47]

Academic assessments of recent experience have been equally skeptical (Martin Feldstein, 1986).

My stylized history has mentioned three ways of using intervention for exchange rate management.

1) It might be used to peg an exchange rate or modify its path. The monetary authorities would purchase all of the foreign or domestic currency that the market was unwilling to absorb. This is sometimes called brute-force intervention.

2) It might be used to alter asset-market equilibrium by changing money supplies or supplies of nonmonetary assets in various currencies. It would be nonsterilized in the first case and sterilized in the second.

3) It might be used to alter expectations by underscoring the authorities' commitment to a particular policy, signalling a future policy change, or making the market more or less confident about its own projections.

We are so accustomed to warning against excessive reliance on intervention that we tend to forget the obvious. A brute-force policy can put the exchange rate wherever the authorities want it, for as long as they are capable of intervening. That is what they did under the Bretton Woods system and what they go on doing under the European Monetary System (EMS). Furthermore, they may actually enhance their ability to peg or move the rate by announcing their intentions. Once they do so, however, their credibility is at stake. That is why governments were so loathe to make exchange rate changes under the Bretton Woods system, when they were committed to pegging rates.

Problems arise when the market comes to believe that the authorities will back off—voluntarily because they have decided that the exchange rate should change, or involuntarily because they are going to run out of reserves. The result is a speculative crisis of the sort modeled by Paul Krugman (1979). The authorities adopt a rigid monetary policy that causes a gradual loss of reserves. Once the nature of that policy is known, because it is announced or readily observable, a speculative attack is inevitable. It will occur as soon as holders of domestic currency realize that the collective effect of their individual actions is certain to exhaust the authorities' reserves or drive them below some minimal level, forcing the authorities to abandon their defense of the exchange rate.

Maurice Obstfeld (1986) has extended Krugman's model to show that there can be a speculative crisis even when it is not foreordained by current policies; it can be produced by self-fulfilling expectations about the policy response to a future crisis. Robert Flood and Peter Garber (1984) have made the money supply wander stochastically around its trend, so that the market cannot predict the path of reserves exactly and the model cannot tell us the date of the crisis. But no one has shown why the authorities should behave so foolishly. If they know what the market knows, that their monetary policy will produce a crisis eventually, why don't they devalue or float the currency before they embark on that policy? To make the crisis model meaningful, we should attach reputational costs to changing the exchange rate and make the authorities weigh them, day by day, against the benefits of continuing to pursue a crisis-inducing monetary policy.

A brute-force policy is usually modeled as continuing flow intervention to keep market forces from changing the exchange rate. Intervention to affect asset-market equilibrium is usually studied by looking at a single act of intervention—an exchange of foreign-currency assets between the central bank and private asset holders.

It is easy to show that a single act of nonsterilized intervention can alter the exchange rate permanently. It affects the fundamentals by changing the money supply. In fact, a nonsterilized purchase of foreign currency has a larger short-run effect on the exchange rate than an open market purchase of domestic securities having the same impact on the money supply (my 1982 paper). By implication, monetary policy should be conducted by purchases and sales of foreign exchange, rather than open market operations in domestic securities, if it is assigned to exchange-rate management, as McKinnon (1986) and Williamson propose. That would buy the biggest bang for a buck.

The effect of sterilized intervention is more controversial. It depends on the degree of substitutability between assets denominated in different currencies. If foreign and domestic bonds were perfect substitutes, sterilized intervention would be futile, because it can merely replace one bond with the other, and this would not matter to asset holders. Even if they are imperfect substitutes, moreover, sterilized intervention has less effect on the exchange rate than an equal amount of nonsterilized intervention (my earlier paper). In some circumstances, however, the authorities may want to alter the exchange rate without affecting the money supply, which raises the crucial quantitative question. Is sterilized intervention powerful enough to be useful?

Simulations by Obstfeld (1983) and others said that it is not, but recent empirical work makes me wonder whether the models they used could capture its whole influence. Wing Woo (1984) has shown that we may understate or miss completely the effects of imperfect substitutability if we do not make sufficient allowance for speculative bubbles. Furthermore, recent attempts to model directly the influence of sterilized intervention say that is may be effective (see, for example Karen Lewis, 1986, and work cited there).

Going back to basics, careful econometric work decisively rejects the joint hypothesis that the forward exchange rate can be taken to predict the rationally expected future spot rate. If expectations were truly rational, we could therefore reject risk neutrality, and bonds denominated in different currencies would not be perfect substitutes. But we have to assimilate two other findings. First, it has been hard to account for the behavior of the risk premium when it is measured in the usual way, by invoking the rational expectations hypothesis. Second, that hypothesis is challenged directly by new work with survey data on exchange rate expectations (Jeffrey Frankel and Kenneth Froot, 1986, and Kathryn Dominguez, 1986). It may be time to measure the risk premium differently —to replace the rational expectations hypothesis with other suppositions about the formation of expectations or to use the survey data. This much seems clear: the use of realized exchange rates to represent expected rates produces a noisy measure of the risk premium (Dominguez and Lewis). In brief, the jury is still out on the size and behavior of the risk premium, and it is not yet possible to decide whether sterilized intervention is a reliable way to alter asset-market equilibrium.

I turn now to the of intervention as a way of changing expectations. Consider first the view most commonly advanced that intervention can be used to let the market know about future policies. This is, of course, a special case of the more general view that governments are justified in trying to alter market prices when they are better informed than the market—in this case, better informed about their own intentions. For the argument to hold in this instance, however, the market must have rational expectations; otherwise, it cannot be expected to draw the appropriate inference about future policies from the exchange rate changes induced by intervention. But other restrictive conditions must hold as well.

First, the authorities must have no other reason for intervening; otherwise, the market cannot know whether the authorities are trying to convey information or trying to achieve some other exchange rate objective. Second, intervention must have a distinct advantage over other ways of conveying information; it must be more persuasive than a simple announcement or a way of avoiding the bureaucratic obstacles to making an announcement. Third, there must be a one-to-one correspondence between the exchange rate change induced by intervention and the future policy it is meant to forecast. This would be true in a simple monetary model, where the exchange rate does not respond to any future policy other than a change in the money supply. It is not true in realistic models. What were the U.S. authorities trying to signal after the Plaza Agreement—a future tightening of fiscal policy or future easing of monetary policy? Because these requirements are so stringent, intervention cannot be a very useful way to provide information about future policies, and I am inclined to emphasize more strongly the other ways of using it to alter expectations.

Intervention can be used for underscoring the authorities' commitment to current policies or for trying to persuade the market that the prevailing exchange rate is inconsistent with the fundamentals—the case of the Plaza Agreement. Whatever the authorities' reason for wanting the market to revise its views, the market is more likely to take heed when the authorities intervene and thus back their words with money. The market will take losses if it bets against them and the authorities prove to be right. Dean Taylor (1982) has tried to show that the authorities have been wrong—that the market has made profits by betting against them—but his calculations have been sharply challenged; they are exceedingly sensitive to the exchange rate chosen for valuing reserves at the end of the period.

Intervention can also be used to change the market's confidence in its own projections, and this may be its most appropriate role. In most theoretical models, especially those in which expectations are rational, all agents hold identical views. In fact, expectations are heterogeneous and held with varying degrees of confidence. This is recognized implicitly in recent work on speculative bubbles, including so-called rational bubbles—instances in which the market has fallen off the saddle path. When expectations are heterogeneous and especially when a bubble appears to be building, intervention may be quite effective. It need not be conducted in brute-force fashion, but rather with the aim of making market participants reassess their views about the likelihood that they will have time to cover their positions before the bubble bursts.

To influence expectations, however, governments must be willing to stand by their views. They should not start to intervene unless they are prepared to persist, which means that they must hold very large reserves or have ample access to short-term credit—a key feature of the EMS that is often ignored but was stressed by the Working Group (para. 52). Furthermore, they should not intervene without saying why. They need not necessarily commit themselves to a target rate or zone. They should be prepared to say that current rates are out of line if they hold that view, and they should not intervene unless they hold that view. They should not lean against the wind unless exchange rates are being blown far off course.

REFERENCES

Dominguez, Kathryn M., "Are Foreign Exchange Forecasts Rational? New Evidence from Survey Data," International Finance Discussion Paper No. 281, Board of Governors of the Federal Reserve System, 1986.

Feldstein, Martin, "New Evidence on the Effects of Exchange Rate Intervention," NBER Working Paper No. 2052, 1986.

Flood, Robert P. and Garber, Peter M., "Collapsing Exchange-Rate Regimes: Some Linear Examples," *Journal of International Economics*, August 1984, *17*, 1–13.

Frankel, Jeffrey A. and Froot, Kenneth, "Three Essays Using Survey Data on Exchange Rate Expectations," Working Paper No. 8614, Department of Economics, University of California-Berkeley, 1986.

Kenen, Peter B., "Effects of Intervention and Sterilization in the Short Run and the Long Run," in R. N. Cooper et al., eds., *The International Monetary System under Flexible Exchange Rate*, Cambridge: Ballinger, 1982.

Krugman, Paul, "A Model of Balance-of-Payments Crises," *Journal of Money, Credit and Banking*, August 1979, *11*, 311–25.

Lewis, Karen K., "Testing for the Effectiveness of Sterilized Foreign Exchange Market Intervention Using a Structural Multi-lateral Asset Market Approach," Working Paper No. 372, Salomon Brothers Center for the Study of Financial Institutions, New York University, 1986.

McKinnon, Ronald I., *An International Standard for Monetary Stabilization*, Policy Analyses in International Economics 8, Washington: Institute for International Economics, 1984.

______, "Monetary and Exchange Rate Policies for International Financial Stability: A Proposal," unpublished, September 1986.

Obstfeld, Maurice, "Exchange Rates, Inflation and the Sterilization Problem: Germany, 1975–1981," *European Economic Review*, March-April 1983, 21, 161–89.

______, "Rational and Self-Fulfilling Balance-of-Payments Crises," *American Economic Review*, March 1986, *76*, 72–81.

Taylor, Dean, "Official Intervention in the Foreign Exchange Market, or, Bet Against the Central Bank," *Journal of Political Economy*, April 1982, *90*, 356–68.

Williamson, John, "Target Zones and the Management of the Dollar," *Brookings Papers on Economic Activity*, 1:1986, 165–74.

Woo, Wing T., "Speculative Bubbles in the Foreign Exchange Markets," Brookings Discussion Papers in International Economics, No. 13, 1984.

Interim Report on Treasury and Federal Reserve Foreign Exchange Operations, August–October 1985, Federal Reserve Bank of New York, *Quarterly Review*, Winter 1985–86.

Plaza Communique, *Group of 5 Statement, September 22, 1985*; reprinted in *IMF Survey*, October 7, 1985.

Working Group, *Report of the Working Group on Exchange Market Intervention*, March, 1983.

Exchange Rate Management: The Role of Target Zones

By JOHN WILLIAMSON*

The essence of the regime of unmanaged floating that prevailed among the major currencies from March 1973 until the Plaza Agreement in 1985 was that the exchange rate was treated as a residual in the process of macroeconomic policy determination. Admittedly there were occasions—such as October 1976 in the case of the pound sterling and October 1978 in the cases of both the U.S. dollar and the Swiss franc—when particular countries became so concerned with a misalignment of their currency that they were forced to abandon "benign (or malign) neglect," but such incidents were episodic. Views about a proper or desirable level of the exchange rate played no systematic role in policy formulation.

Section I explains why I judge the performance of unmanaged floating to have been unsatisfactory. Section II lists the real social benefits that exchange rate flexibility can afford, which should be preserved by any reformed system. Section III describes the target zone proposal and explains why it would preserve the real benefits of flexibility while overcoming the weaknesses of unmanaged floating. Section IV sketches a possible set of comprehensive principles for policy coordination of which target zones would be one natural element.

I. The Failures of Floating

Unmanaged floating has proved unsatisfactory in two key respects. First, it transpired that failure to factor exchange rate implications into the process of policy choice led to recurring, and at times massive, currency misalignments.[1] Misalignments can arise as a rational market response to international differences in real interest rates: in my judgment this explains, for example, much of the initial overvaluation of the dollar in 1981–83, although even here the rise of the dollar during 1982 appears paradoxical. They can also arise as a result of bandwagon effects leading to bubbles in the foreign exchange market: I can, for example, find no other explanation for the continued rise of the dollar from mid-1984 to February 1985, since this was a period when the dollar was already far above any estimate of a sustainable level and interest rate differentials were narrowing (on every plausible basis of measurement).

The second major failure of unmanaged floating is the lack of pressure that it places on countries to coordinate their economic policies. When exchange rates were first allowed to float, most economists regarded the additional independence this afforded economic policy, notably monetary policy, as an advantage. But in retrospect it is far from clear that policy coordination was the irrelevance that this view assumed it to be. The poor performance of the world economy since 1973, including especially the extent of cyclical synchronization and the severity of the debt crisis, is in my judgment partly attributable to the virtual absence of policy coordination.

Admittedly the costs of failing to coordinate policies are still conjectural rather than firmly established. In contrast, the costs of misalignments are glaringly apparent: massive payments imbalances, consequential in-

*Institute for Internal Economics, 11 Dupont Circle, NW, Washington, D.C. 20036. I acknowledge helpful comments on a previous draft from C. Fred Bergsten, Shafiqul Islam, Stephen N. Marris, and Marc Noland. ©Institute for International Economics, all rights reserved; published with permission.

[1] A misalignment is defined as a persistent deviation of the real effective exchange rate from the "fundamental equilibrium exchange rate," the level that can be expected in the medium term to reconcile internal and external balance. These concepts are sketched in Section III.

ternational investment flows that bear no relationship to the real scarcity of capital, distortions to the optimal time pattern of consumption, unnecessary adjustment costs as resources are shifted back and forth between the tradable and nontradable sectors, the destruction of productive capacity, possible ratchet effects on inflation, and protectionist pressures (my 1985 book, pp. 38–45; C. Fred Bergsten, 1986). Indeed, Richard Baldwin and Paul Krugman (1986) argue persuasively that the costs of misalignments have probably been significantly underestimated in the past because of the failure to take account of hysteresis. That is, once a company has abandoned an export market or established itself in an import market, a reversion of the real exchange rate to its initial level will not suffice to restore trade flows to their previous patterns because of the significant overhead costs frequently involved in entering or reentering a market. Adjusting trade flows back after a severe misalignment will therefore tend to be more difficult and costly than traditional econometric estimates suggest.

II. The Social Functions of Exchange Flexibility

The failure of unmanaged floating should not blind one to the fact that fixed exchange rates were abandoned for good reasons. The exchange rate debate has for too long been stilted by excessive emphasis on the textbook cases of fixed and floating rates to the neglect of intermediate regimes, which are motivated by recognition of the weaknesses in both extremes.

A first important function of exchange rate flexibility is that of reconciling differential inflation. Obviously a decision to accommodate inflation through depreciation implies that the exchange rate is not going to be used as a "nominal anchor."[2] There are other and better policies to control inflation: whether anti-inflation policy proves more or less effective than in other countries, appreciation or depreciation is needed respectively to prevent success being undermined by imported inflation or to prevent overvaluation.

A second function is that of facilitating payments adjustment when this proves necessary, by changing the incentives to export and import. It is well known that an exchange rate change is rarely *sufficient* to accomplish adjustment, but, except where disequilibrium is due purely to excess or deficient demand, an attempted adjustment that does not include a change in the exchange rate will involve unnecessarily high unemployment or inflation.

A third function of exchange rate flexibility is that of liberating monetary policy to pursue interest rate targets at variance with those in the rest of the world. If one country is suffering a deeper recession than its partners, it may legitimately wish to ease monetary policy relative to other countries, and that will be feasible only if its currency can depreciate so as to create an expectation of a subsequent rebound that will compensate investors for the temporarily low interest rates. Conversely, a country with abnormally severe inflation may legitimately seek to raise interest rates temporarily, which will require an appreciation. A wide band within which exchange rates are allowed to move around parity provides scope for such temporary variations in monetary policy to pursue anticyclical objectives.

The final legitimate function of exchange rate flexibility is that of absorbing a part of speculative pressures. Instead of requiring that every change in speculative sentiment lead to a change in international reserves and/or interest rates, one can allow changes in the exchange rate to take the strain. Provided these changes do not lead to the prolonged and substantial movements away

[2] I see two decisive objections to the proposal to use a fixed exchange rate as a mechanism for inflation control. The first is that it risks destruction of the tradable goods sector, since the strategy relies on the currency becoming overvalued. The second is that is has a poor track record (compare Britain and France in the 1960's or the Southern Cone in the late 1970's).

from equilibrium that constitute misalignments, they do little harm.

III. The Target Zone Proposal

The target zone proposal envisages a limited number of the major countries negotiating a set of mutually consistent targets for their effective exchange rates. The minimum number of countries needed for a meaningful system would be the three biggest: the United States, Germany, and Japan. Current proposals for policy coordination involve rather more countries: the Group of Five includes the two other countries with currencies in the SDR, namely France and the United Kingdom, while the Group of Seven adds also Canada and Italy.

The aim would be to set exchange rate targets at "fundamental equilibrium exchange rates," that is, at the real values that on average in the medium term are expected to reconcile internal and external balance. This will require agreed interpretations of internal balance (the lowest unemployment rate consistent with the control of inflation) and external balance (a current account balance that is both sustainable and appropriate in the light of thrift and productivity). Both concepts involve an element of subjective judgment and will therefore permit obfuscation by recalcitrant governments, but both provide sufficiently well-defined criteria to form a basis for the sort of technocratic argument that can ultimately lead to international agreement given a modicum of political goodwill. Targets for internal and external balance would then have to be translated into exchange rate targets via some econometric model, which is an essentially technical exercise.

The nominal exchange rate targets corresponding to the agreed real targets should be regularly updated in the light of new data on differential inflation between countries. The real targets should be revised to accommodate both secular trends such as superior productivity growth in the tradable sector (Bela Balassa, 1964) and real shocks or new information.

The participating countries would be expected to conduct their macroeconomic policies with a view to limiting deviations of their exchange rates from the agreed targets, and particularly with a view to preventing exchange rates going outside a broad zone of perhaps ±10 percent around the target. The principal instrument to be used for that purpose would be monetary (interest rate) policy. Provided that the market knew that the authorities were prepared to alter interest rates with a view to managing the exchange rate, there is good reason to believe that jawboning and intervention can also be useful supplementary instruments. If the necessary changes in monetary policy threatened internal balance, it would be necessary to make a compensatory adjustment in fiscal policy.

A country participating in the target zone system need not accept an absolute obligation to keep its exchange rate within the target zone. There are in my view two good reasons for endowing target zones with "soft buffers," which would give a country the right to argue before its peers that it not be required to prevent a breach of the zone. One such circumstance arises where some major shock (such as an oil price change) occurs: rather than forcing the authorities to decide immediately whether to adjust the zone (as might be appropriate if the shock is permanent) or to adjust their policies to push the rate back into the zone, it may be preferable to allow a period for assessment of the magnitude and probable permanence of the shock. This could avoid the danger of countries committing themselves to defense of a disequilibrium rate as used to happen under Bretton Woods.

The second circumstance arises where political cowardice prevents a government taking the fiscal action needed to complement the monetary measures that would be necessary to keep the exchange rate in its target zone. For example, when the dollar first became seriously overvalued in late 1981, the first-best policy would have been monetary relaxation accompanied by fiscal contraction; but had that combination been precluded by political hang-ups, it is arguable that it would have been advisable to maintain monetary discipline in the interests of ensuring success in the battle against inflation even at the cost of the dollar rising temporarily above its target zone.[3] Maintaining the zone under those conditions would nevertheless have warned the market of an official judgment that the rate was overvalued and that policy would in due course have sought a correction, which might at least have avoided the speculative bubble of 1984–85.

A target zone system with these characteristics would in my view provide a viable alternative to both fixed and floating exchange rates, able to limit misalignments and provide a spur to policy coordination (since it would require agreement on target zones, which would in turn need a degree of mutual understanding on policy objectives). It would nevertheless permit exchange rate flexibility to fulfill all four of its genuine social functions: of reconciling differential inflation (by virtue of the provision for automatic adjustment of the nominal target to maintain the target zone constant in real terms); of facilitating payments adjustment (by changing the real zone in response to permanent real shocks); of permitting a degree of independence for anticyclical monetary policy (by virtue of the wide band); and of absorbing speculative shocks (through the wide band and soft buffers).

IV. Policy Coordination

In the spring of 1986, the IMF Interim Committee and subsequently the Tokyo Summit expressed interest in using "indicators" to achieve a more comprehensive framework for policy coordination than that embodied in the target zone proposal. Since I have in part defended that proposal on the basis of the pressure it would create to improve policy coordination, it is natural to complete this paper by laying out my current views on the desirable content of a comprehensive set of rules for coordinated policies.

Incidentally, I do not perceive the choice facing the international community as being one of target zones vs. indicators. The set of rules suggested below embody target zones. Conversely, when I attempted to ask how a presumptive set of rules for policy coordination might be fashioned out of the Tokyo indicators, I ended up with an extended target zone system (my report, 1986).

The rules developed in Hali Edison et al. make use of two intermediate targets, the growth of nominal income and the (real effective) exchange rate. Expressing the internal balance objective in terms of nominal income growth has its disadvantages, notably the lag before nominal income can be observed, but nevertheless appears preferable to alternative specifications such as the Keynesian choice of growth or output (with its danger of accelerating inflation), the monetarist choice of a monetary aggregate (an idea that at one time looked promising but in fact led to fiasco), or the new McKinnon choice of the price level (Ronald McKinnon 1986), which suffers both from the lag problem and from its disregard of the state of the real economy. A target for nominal income growth need not, however, take the naive form of a *constant* growth rate. A sensible formula, which seemed to perform acceptably in our simulations, is to choose a target growth rate of nominal income equal to the sum of the estimated rate of growth of productive potential, plus some fraction of the inherited rate of inflation (to implement a gradualist disinflation strategy), plus a positive function of the deflationary gap.

The exchange rate is a natural intermediate target since the real exchange rate is the dominant medium-run determinant of current account balances apart from income levels, which will presumably bear a reasonably constant relation to capacity in the medium run. (The lag of current balances behind exchange rates is far too long, however, to make it sensible to treat the current balance itself as an intermediate target.)

The assignment rules that I suggest to achieve these intermediate targets are the following:

1) The *average* level of world real interest rates should be revised up (down) if

[3] However, simulations undertaken by Hali Edison, Marcus Miller, and myself (1987) cast some doubt on whether a monetary relaxation inspired by an attempt to keep the dollar in a target zone would in fact have been very damaging to the cause of inflation control.

aggregate growth of nominal income is threatening to exceed (fall short of) the sum of the target growth of nominal income for the participating countries.

2) *Differences* in interest rates among countries should be revised when necessary to limit the deviations of currencies from their target levels.

3) National *fiscal policies* should be revised with a view to achieving national target rates of growth of nominal income.

Rule 1 deals with the $(n-1)$ problem in a McKinnonesque way ("McKinnon without the monetarism"). Rule 2 embodies the essence of the target zone system. Rule 3 endorses Keynesian fiscal policy.

Practical implementation of policy coordination would doubtless be less stark than this summary might suggest. In particular, the guidance to expectations provided by credible target zones plus exchange market intervention plus the wide band will allow significant scope for interest rate differentials to fluctuate with regard to the needs of domestic stabilization. That may well be sufficient to avoid the anticyclical fiscal policy called for by rule 3 requiring the reinstatement of "fine tuning": avoidance of gross mistuning plus the automatic fiscal stabilizers may well suffice. But there is no point in pretending that the world economy can perform satisfactorily irrespective of the fiscal policies pursued by the major powers. Neither can markets be expected to achieve sensibly aligned and reasonably stable exchange rates without the official sector explicitly asking itself what those rates are and being willing to adjust monetary policy to achieve them. Something in the family of target zone ideas is an essential ingredient of any coherent policy coordination program.

REFERENCES

Balassa, Bela, "The Purchasing-Power Parity Doctrine: A Reappraisal," *Journal of Political Economy*, December 1964, *72*, 740–42.

Baldwin, Richard and Krugman, Paul, "Persistent Trade Effects of Large Exchange Rate Shocks," mimeo., 1986.

Bergsten, C. Fred, "Crisis and Reform of the International Monetary System," Ernest Sturc Memorial Lecture delivered at the School of Advanced International Studies, Johns Hopkins University, November 13, 1986.

Edison, Hali, Miller, Marcus and Williamson, John, "On Evaluating and Extending the Target Zone Proposal," *Journal of Policy Modeling*, forthcoming 1987.

McKinnon, Ronald I., "Monetary and Exchange Rate Policies for International Financial Stability: A Proposal," mimeo., 1986.

Williamson, John, *The Exchange Rate System*, Washington: Institute for International Economics, rev. ed., 1985.

______, "Target Zones and Indicators as Instruments for International Economic Policy Coordination," Report to the Group of 24, New York, 1986.

The International Monetary System: Should it be Reformed?

By JACOB A. FRENKEL*

A casual glance through the *Proceedings* of past annual meetings of the American Economic Association reveals that in almost every year during the past twenty years, president-elects of the AEA have devoted at least one session to an examination of issues concerning the international monetary system. Prominent on the agenda has been the question of reform. How should the international monetary system be reformed so as to function more effectively? The premise underlying this question is that the international monetary system has failed and that it must be reformed by an institutional change. In what follows I present some skeptical notes on both the verdict on the failure of the system and on some proposals for reform, especially the target-zones proposal.

To set the stage, it is worth noting that one of the main sources of disenchantments with the present monetary system has been the unpredictability of exchange rates. There has been nothing more confusing than reading through the *ex post* journalistic explanations offered for the day-to-day changes in the U.S. dollar. For example, over the past few years we were told that "The dollar *fell* because the money supply grew faster than expected—thereby generating inflationary expectations," but on another occasion we were told that "The dollar *rose* because the money supply grew faster than expected—thereby generating expectations that the Fed is likely to tighten up and raise interest rates." On another date we were told that "The dollar *fell* since the budget deficit exceeded previous forecasts—thereby generating inflationary expectations on the belief that the Fed will have to monetize the deficit," but, on another occasion we were told that "The dollar *rose* since the budget deficit exceeded previous forecasts—thereby generating expectations that government borrowing needs will drive up interest rates since the Fed is unlikely to give up its firm stance." On yet another day we were told that "The dollar *fell* since oil prices fell—thereby hurting Mexico and other debt-ridden oil-producing countries whose bad fortune may bring about the collapse of important U.S. banks," but, on another occasion we were told that "The dollar *rose* since oil prices fell—thereby helping the debt-ridden oil-consuming countries whose improved fortune will help the vulnerable position of important U.S. banks." More recently the dollar changed again, and this time the explanation was a bit more sophisticated: "The dollar changed because the extent of the revision of the estimated GNP growth rate was smaller than the expected revision of previous forecasts of these estimates." One cannot but sympathize with the difficulties shared by newspaper reporters and financial analysts who feel obligated to come up with daily explanations for daily fluctuations of exchange rates, and one can only imagine the deep frustration that yielded the recent headline in the *International Herald Tribune* according to which "The dollar rose on no news."

The dismal performance of short-term forecasting does not reflect a lack of effort. Rather, it is an intrinsic characteristic of efficient asset markets. Difficulties in forecasting short-term indices of stock markets (like the Dow-Jones index) do not call however, for a reform of the way stock markets operate. For similar reasons one should not assess the performance of the international monetary system on the basis of short-term forecastability of exchange rates. This does

*International Monetary Fund, Washington, D.C. 20431, University of Chicago, and National Bureau of Economic Research. The research reported here is part of the NBER's research programs in International Studies and Economic Fluctuations. Any opinions expressed are my own and not those of the National Bureau of Economic Research.

not imply of course that the present monetary system is without faults or that it should not be reformed. It implies, however, that if a reform is warranted, then it had better be justified on different grounds.

A second noteworthy observation is that over the years, both academics and policymakers have made numerous proposals for reform while, at the same time, the monetary system itself has been in a constant state of change. It evolved from the gold standard to paper money, from the Bretton Woods system to managed float. We also had the Gold Commission but stayed with floating rates, and now attention is focused on target zones, with soft or hard margins.

In spite of the ongoing debate there seems to be little convergence of views about the characteristics of the desired system. This lack of convergence in my view does not reflect lack of effort. Rather, it reflects more fundamental factors that are unlikely to vanish over time. Several are noteworthy. First, participants in the debate have not shared the presumption concerning the relevant alternative to the system they promote. Thus, extreme promoters of fixed rates believe that the relevant choice is between a "good fix" and a "bad flex"; on the other hand, extreme promoters of flexible rates believe that the relevant choice is between a "bad fix" and a "good flex." As is obvious, if these are the alternative choices the outcomes are self-evident, for who would not prefer a "good fix" over a "bad flex"? And, by the same token, who would not prefer a "good flex" over a "bad fix?" In reality, however, the choices are much more complex and much less trivial since they may involve comparisons between a good fix and a good flex or, even more frequently, between a bad fix and a bad flex. When these are the choices, one may expect lack of unanimity. Reasonable people may also differ in their assessments of which "good" system is more likely to gravitate toward its "bad" counterpart. Furthermore, the likelihood that a given good system would deteriorate and be transformed into its bad counterpart depends on the circumstances and, therefore, it is likely that some countries would be wise to choose greater fixity of rates, while other countries would be equally wise to choose greater flexibility.

Second, there are different concepts of the "equilibrium" exchange rate and not all participants in the debate share the same concept. A trivial definition would identify the equilibrium rate as the one that is generated by the free operation of the market place. A more subtle definition emphasizes the *sustainability* of policies as the criterion for equilibrium. Accordingly, if, for example, the current exchange rate reflects unsustainable budget deficits, then this rate is not viewed as an equilibrium rate even though it reflects equality between demand and supply in the market place. An even more subjective view emphasizes the *consequences* of the exchange rate as the ultimate criterion. Accordingly, if the exchange rate yields undesirable results in terms of growth, export, resource allocation, unemployment, and the like, then this rate is not viewed as an equilibrium rate even though it emerges from the market place and reflects sustainable policies.

Third, different countries face different shocks. On purely theoretical grounds, it is clear that the appropriate exchange rate regime depends on the nature and origin of shocks. Are the shocks real or monetary? Are they induced by the private sector or by the public sector? Is their origin domestic or foreign? Are they permanent or transitory? The list of questions is long and circumstances vary across countries and over time.

Fourth, the cost of mistaken policies and the ability to correct errors differ across countries. They depend on the exchange rate regime and on the structural characteristics of the economy. Countries differ from each other in the flexibility of their economic system (for example, the degree of wage indexation, labor mobility, external and internal debt position) as well as in the flexibility of the policymaking process (for example, the speed by which fiscal and monetary policies can be assessed and modified).

Fifth, countries differ from each other according to the various criteria governing the choice of optimal currency areas. These criteria include the degree of openness of the economy, the size of the economy, the degree of commodity diversification, the degree of inflation rates among prospective members, the degree of capital mobility, the degree of other prevailing forms of integration (like custom unions), the degree of similarities of tax structures and other fiscal characteristics, and the degree of similarities of external and domestic monetary and real shocks.

Sixth, views differ about the functions of exchange rates in general and of market mechanisms in particular. On the one hand, there are those who believe that exchange rates are just a nuisance, especially if they move, and anything that moves had better be stopped. (One only wonders whether proponents of this view would also like to see greater fixity of stock market indices?) There are also those who, in spite of the meager evidence, advocate the bubble theory according to which exchange rates have "life of their own" unrelated to "fundamentals." On the other hand, there are those who view exchange rates as an important gauge which provides valuable information about current as well as prospective policies. According to this view, manipulating the exchange rate by intervention and blaming the volatility, unpredictability, and misalignment on the monetary system makes as much sense as blaming the messenger for conveying bad news.

Finally, there are also different views about the advisability and effectiveness of foreign exchange intervention. In spite of growing evidence that the effectiveness of sterilized intervention in exchange rate management is very limited (at least as it operates through the portfolio-balance mechanism), there are those who are still ready to rely on such intervention. In principle, sterilized intervention can be effective by signalling to the market the intent of policymakers. Since the credibility, and thereby the effectiveness, of such signals depend on the track record of past policies, circumstances differ across countries.

The foregoing arguments explain why views about the need for and the desired characteristics of a reform are likely to differ across countries, and may not converge even with the passage of time.

Has the system failed? It is clear that during the past decade foreign-exchange markets have gone through great difficulties. In addition to the volatility and the unpredictability of exchange rates, there is the perception that real exchange rates have been misaligned, and that this misalignment has been costly in terms of resource allocation and general economic performance. The relevant question is whether these faults reflect deficiencies of the *international monetary system* or of *macroeconomic policies*? I believe that faulty policies, especially the lack of synchronization of fiscal policies in the United States, West Germany, and Japan, are at the root cause of the misalignments. Reforming the monetary system without reforming the policies will not do any good and may in fact do harm by diverting the attention from the root-cause of the problem to the monetary system.

There is also the view that the system has failed since it did not yield current-account balance among the major trading partners. Taken by itself, however, this can be viewed as one of the achievements of the monetary system. The ability to rely on international capital markets to smooth out consumption in spite of real shocks may be highly desirable.

We may also wish to ask whether the United States could have carried out its highly successful disinflation policy of the early 1980's while committed to fixed exchange rates? I believe not! The key point that needs emphasis is that the volatility and the misalignment of exchange rates may not be the source of the difficulties, but rather a manifestation of the prevailing package of macroeconomic policies. Fixing or manipulating the rates without introducing a significant change into the conduct of policies may not improve matters at all. It may amount to breaking the thermometer of a patient suffering from high fever instead of providing him with proper medication. The absence of the thermometer will only confuse matters and will reduce the information essential for policymaking. If volatile events and macropolicies are not allowed to be reflected in the foreign-exchange market, they are likely to be transferred to, and reflected in, other markets (such as labor markets) where they cannot be dealt with in as efficient a manner.

The preceding argument ignores, however, one of the important characteristics of the gold-dollar system—the imposition of discipline. Accordingly, it could be argued that the obligation to peg the rate or to follow a predetermined intervention rule would alter fundamentally the conduct of policy by introducing discipline. This view, however, can also be challenged. First, it could equally be argued that by being highly visible flexible exchange rates also impose discipline since current and (expected) future policies are immediately made transparent to both private and public sectors at home and abroad. Indeed, the G–5 Plaza Agreement of September 1985 and the subsequent Paris agreements reached in February 1987 may be viewed as a manifestation of the disciplinary capabilities of flexible exchange rates. Furthermore, it may be argued that national governments are unlikely to adjust the conduct of domestic policies so as to be disciplined by the exchange rate regime. Rather, it is more reasonable to assume that the exchange rate regime is likely to adjust to whatever discipline national governments choose to have. It may be noted in passing that this is indeed one of the more potent arguments against the restoration of the gold standard. If governments were willing to follow policies consistent with the maintenance of a gold standard, then the gold standard itself would not be necessary; if, however, governments were not willing to follow such policies, then the introduction of the gold standard per se would not restore stability since, before long, the standard would have to be abandoned.

Webster's dictionary defines reform as an improvement and a removal of faults. How can anyone be against reform? The key questions, however, are *what* should be reformed, what are the *costs* of the reform and *when* should such reform be adopted. A prerequisite for target zones is that there be agreement on the approximate value of the equilibrium exchange rate, on the boundaries of zones, and on the actions that must take place once the boundaries are reached. At the present such agreement is not in hand. Even if there was agreement on the "equilibrium" exchange rates, one would need to specify in detail what happens if the boundaries are exceeded. It is not enough to say "push them back." We must decide which country should bear the burden of adjustment and which policy will effect that move—monetary, fiscal, government spending, tax? Once this is recognized, it becomes clear that the key difficulties may not lie in the formal structure of the present international monetary system, but rather in the overall mix of macroeconomic policies.

Some say that it is just a matter of tactics whether one examines the system by looking through the exchange rate lens or through the global lens, and that they prefer to focus on the exchange rate lens. I disagree. I believe that the difference between the two lenses is fundamental. It is not a matter of tactics, but is the difference between having a general framework and having a particular framework. It is the difference between patching up a hole here and forgetting that the dam is going to collapse there versus having a consistent set of policies. In principle the adoption of target zones could be acceptable if they encompassed the *entire* array of macroeconomic policies, including in particular fiscal policies. At present the diverging international positions of fiscal policies suggest that it is entirely unlikely that international agreement on such a sweeping reform is feasible. Most of the burden, therefore, is likely to fall on the instruments of monetary policy. As long as fiscal policies are misaligned, a "successful" targeting of the exchange rate by using monetary policies may exacerbate the departures from the optimal mix of fiscal and monetary policies and may be very costly in terms of the overall economic system.

An argument favoring target zones is that the very process of negotiations is likely to enhance the degree of international policy coordination. It must be noted, however, that some successful coordination efforts have also occurred during the past decade (for example, the U.S. dollar support package of November 1978, the Bonn economic summit of 1978, and more recently the G–5 agreement of September 1985). Further, it might be argued that coordination should not be complete, because the perception of independent monetary policy may be necessary for sustaining confidence that monetary policy will not be inflationary in the long run. In addition, there is the danger that the process of negotiating target zones could produce serious frictions among the negotiating parties and could lead ultimately to a reduced level of coordination in this and other areas.

Every system must have a safety valve which allows some flexibility and prevents a crisis and collapse with every conflict. With misaligned fiscal policies and with monetary policies geared towards exchange rate targeting, it would be unfortunate if governments were to exercise their sovereignty by resorting to protectionistic trade policies—to an even greater extent than has been the case under the present system of floating rates with independent monetary policy. The growing frustration with the efforts to reduce the U.S. fiscal deficit by conventional measures have brought about new desperate arguments for the adoption of protectionist measures like import surcharges. The danger with such recommendations is that they might receive the political support of two otherwise unrelated groups. They are likely to gain the support of the traditional advocates of protectionism who claim to defend local industry and workers from what they believe is foreign unfair competition. But, more dangerously, they may gain the support of those whose exclusive concern with the budget deficit leads them to support almost any policy that raises fiscal revenue. Import surcharges, once in place (even those surcharges that are adopted as "temporary measures") are hard to remove since, as George Stigler once remarked, "a sustained policy that has real effects has many good friends." At the present there are very few measures whose long-term costs to the interdependent world economy may be as high as protectionist measures. Taxes on trade will hurt exports, and will restore inward-looking economic isolationism instead of outward-looking economic coordination. Protectionist measures will transmit the wrong signals to those developing countries that are still attempting to resist domestically popular pressures to default on their debt, and, further, they may ignite a trade war. This argument should be considered against the claim that by preventing misalignments of exchange rates target zones reduce the protectionist pressures. With misaligned fiscal policies, the net effect of target zones for exchange rates, implemented through monetary policy, are not clear cut.

The key point made by proponents of target zones is that such a system encompasses the *best* of both worlds—it possess the flexibility of the flexible exchange rate regime as well as the stability of the fixed exchange rate regime. The same logic could be used, however, to argue that this hybrid system encompasses the *worst* of both worlds—it possess the instability of flexible rates and the unsustainability of fixed rates. For in contrast with fixed parties, the target zones are moving. As they move, how do we escape from the inherent difficulty of having the private sector speculate against governments? In the absence of an anchor, what ensures credibility? How exactly are conflicts resolved? What ensures that the moving target zones do not increase turbulence in the foreign exchange market rather than reduce it?

A central feature of any operational monetary system must be a formal resolution of the so-called $n-1$ problem. We have n currencies and only $n-1$ independent exchange rates. We thus have one degree of freedom and its disposal must be explicitly specified. It takes two to tango and it takes one for intervention. The original Bretton Woods system allocated the degree of freedom to the United States which obliged itself to peg the price of gold at $35 an ounce; the other $n-1$ countries then committed themselves to peg their currencies to the U.S. dollar. A design of the international monetary system is not complete unless it provides an explicit resolution to this $n-1$ problem. Therefore, it is essential to ask how the various proposals, including those for target zones, deal with the extra degree of freedom.

As a general rule, a reform of the system should not viewed as an instrument for crisis management. The considerations appropriate for crisis management focus on *short-term* effectiveness. In contrast the considerations appropriate for designing the optimal monetary system should be governed by a *long-term* perspective. The two need not coincide and it is sensible to separate them. In the present context, the short-term crisis concerns the fiscal imbalances in the world economy rather than the monetary system. To be sure, the existing international monetary system is not perfect and it might benefit from a face lift or even from a more drastic reform. A target-zones system is clearly one of the options. But such a reform should perhaps wait until nations restore a more sustainable course of fiscal management.

A reform of the international monetary system should be viewed as a constitutional change that should not be taken lightly. The success of a new monetary arrangement depends on the adoption of a consistent set of policy tools and on a reasonable understanding of the implications of each course of action. In these matters, the cost of delaying the adoption of a new international monetary arrangement until its full implications are understood is likely to be small relative to the cost of a premature implementation. The various proposals for reform of the present international monetary system have many attractions. But since they are novel, prudence is clearly called for. More discussions and critical evaluations can be highly desirable. In view of this it may be a good place to conclude with a quote from John Maynard Keynes' remarks in his closing speech at the original Bretton Woods Conference held over forty years ago. Speaking on the desirability of critical evaluations of the proposed system, Keynes said: "I am greatly encouraged, I confess, by the critical, sceptical and even carping spirit in which our proceedings have been watched and welcomed in the outside world. How much better that our projects should *begin* in disillusion than that they should *end* in it!"

CONTINUING BLACK POVERTY[†]

Earnings Inequality, the Spatial Concentration of Poverty, and the Underclass

By SHELDON DANZIGER AND PETER GOTTSCHALK*

William J. Wilson (1978, 1985, 1986) has hypothesized that the combination of increased spatial concentration and increased inequality of income among blacks has caused adverse behavioral consequences for poor blacks and contributed to the development of an "underclass." Lessening segregation and the general rise in black economic well-being in the postwar period enabled middle-income blacks to move out of segregated inner-city neighborhoods. As a result, low-income blacks in these areas now rarely come in contact with middle-class blacks, who had previously influenced social organizations and community institutions, and provided role models of economic and social success. Wilson hypothesizes that poor blacks have changed their labor force and family behaviors because of the social and economic consequences of this selective outmigration.

Wilson's hypothesis has both an empirical and a causal component. In this paper we focus primarily on the former. In the first two sections, we review changes in the level and distribution of male earnings, and in the spatial concentration of poverty. The trends for blacks are compared to those for whites. The third section discusses the links between the empirical evidence and the causal component—did these changes lead to behavioral responses that contributed to the development of an underclass?

I. Changes in the Level and Distribution of Earnings

Changes in male earnings play an important role in Wilson's theory of the underclass. Rising mean earnings and declining discriminatory practices allow men with higher earnings to move out of inner-city areas. At the same time, declining labor force participation reduces the number of "marriageable men" and increases the number of female-headed families. Thus, Wilson's theory presumes growing inequality among blacks.

We use the computer tapes from the 1940 and 1980 censuses to determine the size of these changes. We examine trends in male labor force participation and in the mean and distribution of annual earnings of black and white males between the ages of 16 and 64, who are not self-employed, nor unpaid family workers, nor in the military or in school.[1]

Between 1939 and 1979, there was a large increase in the proportion of black males who reported zero annual earnings—from 14.5 to 21.2 percent. For whites, the proportion declined from 12.6 to 9.6 percent. The increase in blacks without earnings may affect the mean earnings of blacks relative to

[†] *Discussants*: Barbara A. P. Jones, Clark College; Gilbert Moore, Morehouse College.

*Director, Institute for Research on Poverty, University of Wisconsin, Madison, WI 53706, and Professor of Economics, Bowdoin College, Brunswick, ME 04011, respectively. This research was supported in part by a grant from the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation to the Institute for Research on Poverty. Steven Berry, Hamid Davoodi, and George Slotsve provided research assistance; Cathy Esser, clerical assistance; Elizabeth Uhr, editorial assistance. Any opinions expressed are ours alone.

[1] The self-employed were excluded because their earnings were not counted in the 1940 census.

whites, for, if blacks with below-average earnings drop out of the labor market, the mean of those remaining will increase even if blacks with earnings do not experience increases (see Charles Brown, 1984; Richard Butler and James Heckman, 1977; William Darity and Samuel Myers, 1980; and Wayne Vroman, 1986). Likewise, inequality of earnings may decrease because of this selection.

To account for selection, we present data both for all males in our sample and for those with earnings. These data reflect two extreme assumptions. If the earnings opportunities of those not working are the same as of those currently working, then an analysis of earners gives unbiased estimates of the parameters of the earnings distribution. If, however, there are no jobs for those without earnings, then the "zeroes" represent true opportunities and should be included. Accounting for selection would provide an intermediate position by imputing nonzero earnings opportunities to nonearners.

Table 1 presents the earnings' share in 1939 and 1979 of each quintile for all males and for those with earnings. For whites, the changes for all males are similar to those for earners—the shares of the bottom three quintiles increased, while those of the top two decreased. The trends for blacks are quite different. For all males, the shares of the bottom two quintiles declined substantially, those of the next two increased, and that of the top quintile declined marginally. For those with earnings, the shares of the bottom and top quintiles declined.

Table 2 documents changes between 1939 and 1979 in mean earnings and three summary measures of inequality—the coefficient of variation, the Gini coefficient, and the variance of the log of earnings. Mean real earnings increased substantially for both blacks and whites, with blacks gaining relative to whites—the black-white ratio of means rose from .44 to .60 if all males are included, and from .43 to .69 if only earners are included.

The results are similar for the distributional measures. The increase in the black-white ratio for each summary measure of inequality shows that black inequality increased relative to that of whites over the

TABLE 1–SHARE OF AGGREGATE ANNUAL EARNINGS RECEIVED BY EACH QUINTILE OF MALES AND OF MALES WITH EARNINGS, 1939 AND 1979[a]

	Blacks		Whites	
	1939	1979	1939	1979
All Males[b]				
1	.77	.00	.93	1.92
2	7.97	5.44	8.21	11.78
3	15.72	17.49	16.77	18.86
4	26.16	28.30	26.18	25.76
5	49.41	48.78	47.90	41.67
Males with Earnings				
1	4.51	3.61	4.02	5.30
2	9.84	11.58	10.40	12.98
3	16.42	17.93	17.17	18.48
4	24.99	25.77	24.78	24.37
5	44.74	41.11	43.63	38.87

Source: For Tables 1 and 2, computations by authors from computer tapes of 1940 and 1980 Censuses of Population.

[a] Shown in percent.

[b] All males includes those between the ages of 16 and 64, who are not self-employed nor unpaid family workers, nor in the military or in school.

TABLE 2—THE LEVEL AND DISTRIBUTION OF MALE EARNINGS, 1939 AND 1979[a]

	Blacks (1)	Whites (2)	Ratio (1)/(2)
1939			
All Males			
Mean	$2461	$5582	0.44
Coeff. var.	.917	.890	1.03
Gini coeff.	.492	.474	1.04
Var. ln	1.670	2.246	0.74
Males with Earnings			
Mean	$2672	$6261	0.43
Coeff. var.	.799	.776	1.03
Gini coeff.	.405	.399	1.02
Var. ln	.770	.842	0.91
1979			
All Males			
Mean	$9276	$15,404	0.60
Coeff. var.	.937	.747	1.25
Gini coeff.	.507	.397	1.28
Var. ln	3.996	2.652	1.51
Males with Earnings			
Mean	$11,592	$16,881	0.69
Coeff. var.	.708	.647	1.09
Gini coeff.	.378	.335	1.13
Var. ln	1.082	.764	1.42

[a] To compute the log variance, $100 in 1979 dollars was added to each observation. All figures shown are in 1979 constant dollars.

forty-year period. However, the ratio increases more rapidly when all males are included. Thus, studies that concentrate only on earners (for example, James Smith and Finis Welch, 1986) will show greater black progress relative to that of whites, and smaller growth in inequality among blacks than studies that include all persons.

Because of the particular changes in the shape of the distributions, the trends shown by the three summary measures vary dramatically. The coefficient of variation and the Gini coefficient for all black men increase moderately between 1939 and 1979. However, the variance of the log of earnings, which disproportionately weights changes at the bottom of the distribution, more than doubles. This reflects the declining shares of the bottom two quintiles shown in Table 1. The increases in mean earnings and in inequality are consistent with Wilson's hypothesis.

II. Changes in the Spatial Concentration of Poverty

We use published Census Bureau data (not available prior to 1969) on the number of poor persons living in urban poverty areas to measure the spatial concentration of poverty. These areas are defined as clusters of census tracts in which at least 40 percent of the residents are poor. By calculating the change between 1969 and 1979 in the number of people living in such areas in the fifty cities that had the largest population in 1970, we can get a rough measure of recent changes in spatial concentration.[2] However, Wilson's hypothesis, which presumes both selective outmigration and increasing poverty within poverty areas, cannot be fully tested because the number of people in poverty areas can grow for several different reasons. First, those tracts defined as poverty areas in both 1969 and 1979 may include more poor persons either because more poor people moved into them, or because a greater percentage of existing tract residents became poor. Furthermore, tracts which were not in poverty areas in 1969 could have been included by 1979 either because nonpoor residents moved away or because the number of poor within them increased. Thus, one can determine if concentration has increased, but not whether it is due to selective outmigration or increased poverty within these areas.

TABLE 3—NUMBER OF PERSONS AND POOR PERSONS LIVING IN THE UNITED STATES, THE 50 LARGEST CITIES, AND POVERTY AREAS IN THESE CITIES, BLACK AND NONBLACK, 1969 AND 1979[a]

Populations	Black	Nonblack	All Persons
A. All Income Levels			
United States			
1. 1969	22,034	177,528	199,562
2. 1979	25,967	196,869	222,838
3. % Change	+17.8	+10.9	+11.7
50 Largest Cities			
4. 1969	9,874	29,954	39,828
5. 1979	10,588	27,242	37,830
6. % Change	+7.2	−9.1	−5.0
Poverty Areas in 50 Largest Cities			
7. 1969	1,412	506	1,918
8. 1979	2,183	1,038	3,221
9. % Change	+54.6	+105.1	+67.9
B. Poor Persons			
United States			
10. 1969	7,095	17,107	24,147
11. 1979	8,050	18,022	26,072
12. % Change	+13.5	+5.3	+7.9
50 Largest Cities			
13. 1969	2,692	3,312	6,004
14. 1979	3,140	3,568	6,708
15. % Change	+16.6	+7.7	+11.7
Poverty Areas in 50 Largest Cities			
16. 1969	708	266	974
17. 1979	1,124	490	1,614
18. % Change	+58.7	+84.2	+65.7

Source: Computations from matching the 50 SMSAs in 1970 Census of Population Subject Reports, *Low Income Areas in Large Cities*, Table 1, with the same SMSAs in 1980 Census of Population Subject Reports, *Poverty Areas in Large Cities*, Table 1.

[a] Poverty areas are defined as clusters of census tracts with a poverty rate of 40 percent or more. All figures are shown in thousands.

Panel A in Table 3 shows the population of the United States in 1969 and 1979, and the percentage change over the decade. The same data are given for all persons living in

[2] This methodology was brought to our attention by David Ellwood.

the fifty cities with the largest populations in 1969, and for the number living in poverty areas in these cities. Panel B of Table 3 gives the same data for poor persons.

While our emphasis is on the increased spatial concentration of poverty, note that persons living in poverty areas comprise a relatively small group. In 1979, the 3.221 million people living in urban poverty areas made up only 1.4 percent of the U.S. population; the 1.614 million poor persons living in these areas made up only 6.2 percent of all poor persons.[3]

The spatial concentration of the poor increased over this decade. The number of poor persons living in poverty areas in the fifty largest cities increased more rapidly than the number of poor persons in these cities or in the United States (compare row 18 to rows 12 and 15). Poor blacks living in poverty areas as a percentage of poor blacks in the United States grew from 10.0 to 14.0 percent. For poor nonblacks, the proportion increased from 1.6 to 2.7 percent.[4] The spatial concentration of the poor within the largest cities also increased between 1969 and 1979—the proportion of the urban poor living in poverty areas increased from 26.3 to 35.8 percent for blacks and from 8.0 to 13.7 percent for nonblacks.[5] Therefore, Wilson is also correct in claiming an increased concentration of urban poverty.

III. Testable Implications of Wilson's Hypothesis

Geographic concentration of the poor is undesirable on its own account, as low-income people living in poor areas have fewer public amenities than those who live in other areas. In addition, Wilson argues that increased male earnings inequality, reduced male labor force participation, and desegregation have all contributed to polarization within the black community and a cycle of economic concentration and decline. As successful blacks moved away, social organization and community institutions deteriorated, opportunities declined, and the attitudes of those who remained changed. These changes, by further weakening social norms and community support structures, decreased the probability that those who remained could escape poverty. They also reduced intergenerational mobility as young ghetto residents came to perceive high rates of unemployment, illegal activities, welfare receipt, and out-of-wedlock births as the status quo. The resulting behavioral changes increased the incentive for the remaining nonpoor residents to move away, further reducing the probabilities of escaping poverty for those who remained.

While this thesis has intuitive appeal, much of it has not yet been tested and some of it may never be tested.[6] If increased inequality and high concentrations of the poor provide fertile ground for the development of an underclass, then the census data have merely shown that the ground became more fertile. A full test of the hypothesis requires evidence on several separate questions. Has the behavior among poor blacks changed, and, if so, have these changes been largest in areas where black middle-class flight has been the largest? Since many of the model's predictions imply reduced chances for children to escape poverty, longitudinal data (which is not yet be available) are needed.

Further, even if measured behavior in poverty areas could be shown to have changed (for example, labor force participation, illegitimacy, and crime rates), one must show that this change was caused by the selective outmigration. An alternative explanation is that the observed change in behavior simply resulted from changes in the composition of the community. Suppose that people have always differed in their attachment to work and that those most prone to

[3] Blacks living in poverty areas comprised 8.4 percent of all blacks and 14 percent of poor blacks in 1979.

[4] These percentages can be computed from the bottom panel of Table 3 as follows: for 1960, the ratio of (row 16/row 10); for 1979, the ratio of (row 17/row 11).

[5] From the bottom panel of Table 3, the 1969 percentage is the ratio of (row 16/row 13), and the 1979 percentage, the ratio of (row 17/row 14).

[6] Wilson is currently supervising an intensive study of poverty and the underclass in Chicago.

work were the first to leave urban ghettos. The result would be that the average labor force attachment of the remaining ghetto residents would decline, even if those left behind did not change their behavior at all in response to the selective outmigration. A similar argument would apply to other behaviors, such as the out-of-wedlock birth rate or the crime rate. Wilson hypothesizes that in addition to this selectivity explanation for the change in behavior within the poverty area, declining social supports and economic opportunities contributed to an adverse response among the remaining residents that further lowered the mean behavior.

A final question would remain even if one could show that the behavior of individuals left behind had changed *and* that it was due to changed attitudes. One would still have to show that the attitude changes were due to the selective outmigration from these areas and not from other possible explanations, such as increased permissiveness in the society at large or increased work disincentives in transfer programs.

In conclusion, despite real earnings gains that exceed those of whites on average, black male earnings inequality and labor force withdrawal have increased. And, the number of poor blacks living in poverty areas has increased in absolute numbers and as a percentage of all of the poor. These economic trends are consistent with Wilson's hypothesis, but whether they have contributed to the development of an underclass among the black poor is a question which cannot yet be answered.

REFERENCES

Brown, Charles, "Black/White Earnings Ratios since the Civil Rights Act of 1964: The Importance of Labor Market Dropouts," *Quarterly Journal of Economics*, February 1984, *99*, 31–44.

Butler, Richard and Heckman, James J., "The Government's Impact on the Labor Market Status of Black Americans: A Critical Review," in Leonard J. Hausman et al., eds., *Equal Rights and Industrial Relations*, Madison: Industrial Relations Research Association, 1977, 235–81.

Darity, William, Jr. and Myers, Samuel, Jr., "Changes in Black-White Income Inequality, 1968-1978: A Decade of Progress?," *Review of Black Political Economy*, Summer 1980, *10*, 355–79.

Smith, James and Welch, Finis, *Closing the Gap: Forty Years of Economic Progress for Blacks*, Santa Monica: Rand Corporation, 1986.

Vroman, Wayne, "The Relative Earnings of Black Men: An Analysis of the Sample Selection Hypothesis," mimeo., Urban Institute, 1986.

Wilson, William J., *The Declining Significance of Race*, Chicago: University of Chicago Press, 1978.

______, "Cycles of Deprivation and the Underclass Debate," *Social Service Review*, December 1985, *59*, 541–59.

______, "Social Policy and Minority Groups," Institute for Research on Poverty Conference Paper, University of Wisconsin-Madison, 1986.

Do Transfer Payments Keep the Poor in Poverty?

By WILLIAM A. DARITY, JR. AND SAMUEL L. MYERS, JR.*

Do transfer payments keep the poor in poverty? There are at least two ways to restate the question, two ways that imply quite different, although related, lines of inquiry. A first restatement is: Do transfer payments raise the poor out of poverty? In short, are transfer payments efficacious in achieving the antipoverty function many of them obstensibly were introduced to perform, particularly those that are means-tested income redistribution programs? A second and more provocative restatement is: Do transfer payments cause the nonpoor to enter poverty, or, only a bit more mildly, do transfer payments induce the poor to stay in poverty? In short, does the existence of a social welfare system of the type extant in the United States have the perverse effect of increasing the incidence of poverty or perpetuating the poverty status of those who already are poor? In this paper we examine both lines of inquiry. We suspect that the net consequence of our investigation will be to raise troubling questions about the structure of American society and the fundamental causes of poverty.

I. Antipoverty Effectiveness of Transfer Programs?

Plainly, federal income-support programs reduce the numbers of individuals, households, or families whose income falls below the officially designated poverty line. Given the official criterion, the percentage of the population that is counted as poor is smaller after income transfers are taken into account. In 1967, for example, more than one-quarter of the pre-transfer poor were pushed above the poverty line after transfers; in 1978, 43.6 percent of the pre-transfer poor were pushed above the poverty line after receipt of transfers. By 1982 the proportion lifted above the Orshansky line had declined slightly to 37.5 percent. Nevertheless, both post-transfer and pre-transfer poverty rates in 1982 exceeded those in 1967 (Sheldon Danziger and Daniel Feaster, 1985, pp. 91–92). In fact, pre-transfer poverty rates have risen secularly since 1969 (Danziger and Peter Gottschalk, 1985, p. 34). Danziger and Feaster report that the decline in the relative antipoverty effectiveness of income-support programs—both cash and in-kind, and between-kind—between 1978 and 1982 was due primarily "to the declining value of transfers and not to macroeconomic conditions" (p. 102). Substantial funding reductions involving both loss of coverage as well as smaller amounts for those still eligible in the first year of the Reagan Administration took place in means-tested welfare programs (Danziger and Feaster, p. 109), so that the targeted programs became less effectual in reducing the incidence of poverty.

In a separate study, Danziger (1983) also took a brief comparative look at the incidence of pre-transfer and post-transfer poverty rates by race by head of household using 1980 data from the *Current Population Survey* after cash transfers. In every age-sex category, the proportionate impact of cash transfers on the occurrence of poverty was larger for whites than nonwhites. This holds true even for the elderly, the group reaping the largest reduction in poverty rates after redistribution. Nonaged, nonwhite female heads of households experienced the smallest reduction in poverty incidence after receipt of cash transfers of any of the groups considered by Danziger, a striking finding given the significant presence of female-headed households among the nonwhite poor.

We also have examined (1986) pre-and post-transfer poverty rates using data from the 1980 decennial census with detailed breakdowns of the impacts by race and

*Department of Economics, University of North Carolina, Chapel Hill, NC 27514, and Afro-American Studies Program and Department of Economics, University of Maryland, College Park, MD 20742.

ethnicity of household head. There we confirmed the existence of major racial differences in the effectiveness of transfers in reducing poverty. An updated analysis of the census data supplied to us by Danziger reinforces this finding further. This analysis focuses on family heads who are also heads of households. Two transfer programs are emphasized: Public Assistance and Social Security. Given a poverty line of $7,412 for a family of four in 1979, the proportionate impact on poverty rates of these two transfers was greatest for white male-headed families. These families, with pre-transfer poverty rates of 10.4 percent, realized a huge 55.8 percentage reduction in poverty when transfers lowered their poverty rates to 4.6 percent.

By comparison, female-headed families among blacks and American Indians in reservation states—families whose nearly identical post-transfer poverty rates of 45 percent are nine points lower than their pre-transfer rates—only realized 16 to 17 percentage reductions in their poverty rates as a result of receipt of Social Security or public assistance benefits. Even white female-headed families, with poverty rates of 30.4 and 20.8 percent before and after transfers, had larger proportionate reductions in poverty.

Are these racial differences in the proportionate reductions in poverty simply an artifact of the higher initial poverty rates of minorities? No. The *absolute* reductions in poverty from transfers are in fact smaller for many minority groups than they are for whites. For example, among all male-headed American Indian families, transfers reduced poverty rates from 20.4 to 15.5 percent. This reduction is *smaller* than the absolute reduction in poverty among white male-headed families.

Are these racial differences a quirk that arises from ignoring unrelated individuals in measuring family poverty rates? No. When households, including single-person households, are examined, the findings still hold. Poverty rates are generally higher among households than families. Among minorities, however, they are not much higher in households headed by women than they are in families. But among whites, pre-transfer poverty rates are far higher for female-headed households than they are for families. Since the post-transfer poverty rates are about the same for both households and families among whites, the racial gap in the poverty-reducing effectiveness of transfers becomes even more pronounced.

What determines these racial differences in the antipoverty effectiveness of transfers is far from clear. What is clear is that welfare and Social Security do not have uniform poverty reduction impacts across various racial and ethnic groups in America.

II. Transfer-Induced Poverty?

There is, however, an artificial feature to the calculations performed above that compare pre-and post-transfer poverty rates. Households do not, in fact, experience the pre-transfer levels of income. Eligibility is dictated by past income net of means-tested income received by the individual or household in question. Therefore, pre-transfer income, strictly speaking, is a hypothetical construct arrived at by deducting all transfer payments from estimates of actual household incomes. Moreover, individuals and households presumably know that such transfer programs are available to them (although Richard Coe has demonstrated that "more than 40 percent of the eligible nonparticipants in the food stamp program did not believe they were eligible to participate," 1983, p. 1051). It would seem reasonable to suppose that their actions and decisions would be considerably different if the scheme of transfer payments did not exist, particularly their actions and decisions that affect the generation of personal and household income. Therefore, arguably their incomes (and the overall distribution of income) in a world without recourse to income supports would look very different from the incomes (and the overall distribution of income) hypothetically ascribed to them by subtracting transfers they actually received from their total incomes.

Some critics of the welfare state in general and the Great Society programs in particular contend that conditions would be so different that fewer of the poor would still be with us in the absence of transfer payments. They claim that the existence of income supports destroys motivation and induces a certain indolence and idleness among significant numbers of the American population—especially those most likely to gain initial entry into the labor market in relatively low-paid occupations. If there were no income transfer programs—or if their magnitude was considerably smaller—there would be, these critics charge, significantly fewer poor people than the estimates of pretransfer poor in the previous section suggest. Charles Murray makes the point bluntly: "We tried to provide more for the poor and produced more poor instead. We tried to remove the barriers to escape from poverty, and inadvertently built a trap" (1984, p. 9).

Transfer payments clearly lift many of the poor above the poverty line and close the poverty gap for many others. But Murray's position suggests that there are more pretransfer poor to be assisted by redistributive measures because the transfer programs are available. He points to the finding that the estimated proportions of pre-transfer poor have grown since the late 1960's as proof of his case; he attributes the growth to the disincentive effects (or perverse incentive effects)—particularly on work effort—he identifies in transfer payments.

It is probably more useful to evaluate the specific mechanisms that have been advanced as ways in which transfer payments can produce poverty, rather than alleviate poverty. These mechanisms can be summarized in three major categories: 1) negative labor supply effects; 2) negative effects on family structure—both in terms of composition and intergenerational replication of family type; and 3) adverse psychological and material dependency effects. We consider each of these in turn.

Virtually all researchers who have investigated the effects of the existing scheme of transfer payments on labor supply have found statistically significant net negative impacts. Simultaneously virtually all researchers have found the impacts to be small. In our own research (1980, p. 372), we also found small effects and, intriguingly, a larger negative labor supply effect for whites than blacks. In a survey article on the various effects of income transfer programs, Danziger et al. (1982) inferred from the research available that existing transfer programs reduced aggregate labor supply by 4.8 percent. David Blau and Philip Robins estimated on the basis of the same survey article "that welfare recipients reduce hours of work about 1% relative to total work hours of all workers" (1986, p. 83).

In the study providing perhaps the strongest case for significant work disincentives, Blau and Robins (p. 94) find that welfare recipients are less likely to enter employment and are more likely to exit from employment than nonwelfare recipients. However, *after* adjusting for individuals' personal characteristics aside from welfare-nonwelfare status, they find that "youths in welfare families are less likely to leave employment than youths in nonwelfare families" (pp. 100–01). The same is true for men and married women receiving welfare, "although the effects are not statistically significant." The results remain somewhat mixed because "[t]these are the only cases in which the welfare effects on transitions into and out of employment changes sign when compared to the unadjusted differences." Blau and Robins (p. 94) also find that unemployed welfare recipients are more likely to stay in the labor force than nonwelfare recipients, but they (p. 99) speculate that this may be due to legal job search requirements to maintain eligibility. Nevertheless, Blau and Robins do not contend that there are aggregate reductions in labor supply in excess of those reported by Danziger, Robert Havemen, and Robert Plotnick (1981).

Implicit in this line of inquiry is the notion that if there are major adverse labor supply effects from transfer payments that the critics of social programs actually may have a case. But the demonstration that there are negative labor supply effects, whether large or small, does not really address the question of whether or not transfer payments produce poverty. If welfare recipients, for example, were to supply more labor would they be nonpoor? What types of occupations and what wages are they likely to obtain?

Suppose further that today's welfare recipients were to supply more labor in an environment where transfer payments did not exist altogether. One can suppose that the abolition of transfer payments—or a far more drastic reduction than that enacted at the start of the Reagan Administration—would have repercussions on the contour of wages. Presumably drastic reductions in money wages would follow from the removal of income supports. If commodity prices do not follow money wages on their downward path it is a recipe for reductions in the real earnings of substantial segments of the *employed* workforce. Conceivably poverty would expand significantly. It is also quite conceivable that a commodity price deflation would follow on the heels of the money wage decline (see J. M. Keynes, 1936, ch. 19, and Amitava Dutt, 1986–87, pp. 283–88). This would be no more of a felicitous outcome for the economy, particularly given its preexisting structure of consumer and corporate indebtness. Now we would have the recipe for a classic Fisherian debt-deflation, with its cyclically impoverishing implication.

Research that has focused on the labor supply effects of transfer payments implicity treats labor market conditions as unchanged if the transfer payments no longer were available, patently an unrealistic assumption.

A dramatic reduction in transfer payments would trigger a depressant shock wave along the structure of money wages. An increase in transfer payments that genuinely eliminates poverty would place in jeopardy the willingness of workers to take a host of low and moderately paid jobs. It is in the first respect that it cannot be said that elimination of transfer payments will eliminate poverty. It is in the second respect that the current system of transfer payments is not designed to eliminate poverty.

Murray and Richard Vedder and Lowell Gallaway (1986) also claim that the nature of income support programs, especially AFDC, promote teen pregnancy and divorce, both proximate causes of the growth in female-headed families. Female headship, in its turn, correlates strongly with poverty status. We examined this proposition in detail (1983 and 1984) and found no correspondence between the magnitude of transfer payments and growth in female-headed families among blacks using post-1950 time-series data. Fertility and family formation among low-income black women did not prove to be terribly sensitive to variations in the magnitude of transfer payments. Of course, our use of standard regression techniques did not enable us to address causal processes that feature cumulative, sleeper, and/or threshold effects. However, we did identify other factors that contribute in an important fashion to the increase in female headship among blacks. Of special significance was the low male-female ratio among blacks, indicative of the marginal status of black men in modern America. Marginal men lead to families living on the fringes, with women bearing an immense burden of rearing new generations with few resources —financial, communal, or spiritual (also see James Stewart and Joseph Scott, 1978).

Some cross-section studies (for example, Danziger et al.) do find correlations between variations in AFDC payments and the incident of female headship. But these studies uniformly fail to separate the formation of female-headed families from the tendency of female-headed families—once formed—to cluster in high payments states via migration. There is solid evidence to suggest that public welfare program recipients do migrate to states where they can receive the best package of benefits (see Laurence Southwick, 1981, and Edward Gramlich and Deborah Laren, 1984). Moreover, evidence on the family structure patterns of native Americans seems to reinforce the emphasis we have given to the availability of males as mates. Native Americans have comparable, if not higher, levels of welfare recipiency as black Americans but not the same incidence of female headship. Female-headed families are closer to one-quarter of American Indian families, unlike the close to 50 percent figure for black Americans. Gary

Sandefur and Trudy McKinnell (1987) suggest that this differential may be due to the relatively small size of the native American population (less than 1 percent of the U.S. population) and the concomitantly greater ease of intermarriage with non-Indians.

Another variant of the argument concerning transfer payments effects on family structure and hence on economic status is the claim that there is a welfare culture of families headed by women that now tends to reproduce themselves from generation to generation. Racial differences in exposure to welfare programs are pronounced. As Martha Hill and Michael Ponza report: "Both welfare receipt and especially high levels of welfare dependency were more common among black children than among white children. The majority (about 60 percent) of black children grew up in families receiving welfare at some time, whereas one-fifth of white children grew up in welfare recipient families" (1983, pp. 20–21). But does exposure to welfare lead an individual to be on welfare as well when they reach adulthood? In particular, is this the case for women?

Sara McLanahan (1985) has found that (a) growing up in a female-headed family increases the likelihood significantly that a woman will form a female-headed family of her own as a young adult or as a mature adult, and (b) if her childhood family received welfare, it raises the probability that her own family will be a welfare-receiving family as well. However, McLanahan also finds, paradoxically, that "although being on welfare during adolescence is associated with being on welfare in adulthood, the relationship is weaker the longer the family received welfare" (p. 24). Therefore, "...the most dependent families (those on welfare for all five adolescent years) are *not* the most likely to reproduce female-headed families which encounter to what is predicted by most theories of intergenerational dependency" (p. 24).

III. Losing Ground?

But the substantive issue is the following: does the exposure to welfare predispose the young women to form welfare families, or do they receive welfare both as children and as adults because they are condemned to poverty by other forces? Murray (1986) himself has suggested that blacks have such a high incidence of female headship and welfare recipiency because such a large portion of the black population is concentrated in the underclass. He points out that if one looks at the proportionately smaller white underclass, the same family structure and welfare dependency patterns, broadly speaking, obtain. To make his point, Murray examines data from the heartland of America, data from the state of Ohio. But Murray then backs off from this profound insight to argue that "the growth in the underclass is largely the direct, not so mysterious result of bad social policy" (1986, p. 34). This, of course, neglects to explain why the underclass exists in the first place—aside from its growth–nor does it explain why the black underclass is so much larger than the white underclass. Why would social policy exercise such sharply differentiated effects unless there were historic differentials that existed prior to and independently of the Great Society welfare programs? There may be truth in Murray's claims that social policy has negative impacts, but it is not the case that elimination of those policies will, in and of itself, produce a healing environment.

Where the transfer-induced poverty argument has greatest force in our estimation is with respect to the charge that particular ethnic and racial groups are rendered dependent on the public largess. These groups are necessarily subject to the redistributive "goodwill" of the rest of the population. If that goodwill is withdrawn, they have no independent basis for community support. This is precisely what Gary Anders (1981) sees in his analysis of the case of the Cherokee people as an instance of the general phenomenon of the underdevelopment of the American Indian people, what he describes as a reduction from self-sufficiency to poverty and welfare dependence.

Michael Brown and Steven Erie see a similar entrapment occurring for black Americans: "Ironically, federal social welfare policy may be performing a control as well as a reward function for blacks by restructuring the economic relationship of the middle class to poor and by generating incentives to maintain the new class relationship. The social welfare economy tends to reinforce a dual labor market, particularly in the black community" (1981, pp. 329–30). Social policy (i.e., the welfare state) induces a restructuring of intragroup class cleavages, placing publicly funded professional social service providers on one side, and an underclass of social service recipients on the other side. Most perverse, the occupational status of the middle class is contingent on the continued maintenance of the underclass as an underclass and the continued existence of the prevailing scheme of social welfare programs, regardless of their effectiveness.

A similar pattern of events increasingly is apparent among native Americans as well with the growth of the tribal social service bureaucracies administered by the Indian middle-class using public funds (see David Russakoff, 1983).

In this light, the antipoverty impact of transfer programs may be of minor importance when weighed against the group vulnerability and class cleavage effects they propagate. To fall under the way of the existing scheme of social welfare programs may mean that a people sacrifice their ability to shape their own future. The trap is not merely pecuniary in the narrow sense that Murray conceives it, the trap of poverty. It is the potential trap of a loss of any automomous capacity to shape the destiny of one's own people—a fairly steep price to pay for crossing the Orshansky line.

REFERENCES

Anders, Gary C., "The Reduction of a Self-Sufficient People to Poverty and Welfare Dependence: An Analysis of the Causes of Cherokke Indian Underdevelopment," *American Journal of Economics and Sociology*, July 1981, *40*, 225–37.

Blau, David M. and Robins, Philip K., "Labor Supply Response to Welfare Programs: A Dynamic Analysis," *Journal of Labor Economics*, January 1986, *4*, 82–104.

Brown, Michael K. and Erie, Steven P., "Blacks and the Legacy of the Great Society: The Economic and Political Impact of Federal Social Policy," *Public Policy*, Summer 1981, *29*, 229–330.

Coe, Richard D., "Nonparticipation in Welfare Programs By Eligible Households: The Case of the Food Stamp Program," *Journal of Economic Issues*, December 1983, *17*, 1035–74.

Danziger, Sheldon, "Budget Cuts as Welfare Reform," *American Economic Review Proceedings*, May 1983, *73*, 65–70.

______ **and Feaster, Daniel,** "Income Transfers and Poverty in the 1980s," in John Quigley and Daniel Rubinfield, eds., *American Domestic Priorities*, Berkeley: University of California Press, 1985.

______ **and Gottschalk, Peter,** "The Poverty of *Losing Ground*," *Challenge*, May/June 1985, *28*, 32–38.

______, **Haveman, Robert and Plotnick, Robert,** "How Income Transfers Affect Work, Savings, and the Income Distribution: A Critical Review," *Journal of Economic Literature*, September 1981, *19*, 975–1028.

______ **et al.,** "Work and Welfare as Determinants of Female Poverty and Headship," *Quarterly Journal of Economics*, August 1982, *97*, 519–34.

Darity, William A., Jr. and Myers, Samuel L., Jr., "Changes in Black-White Income Inequality, 1968–1978: A Decade of Progress?," *Review of Black Political Economy*, Summer 1980, *10*, 354, 356–79.

______ **and** ______, "Changes in Black Family Structure: Implications for Welfare Dependency," *American Economic Review Proceedings*, May 1983, *73*, 59–64.

______ **and** ______, "Does Welfare Cause Female Headship? The Case of the Black Family," *Journal of Marriage and the Family*, November 1984, *46*, 765–79.

______ **and** ______, "Transfer Programs and the Economic Well-Being of Minorities," prepared for the Institute for Research on Poverty Conference on Poverty and Social Policy: The Minority Experience, November 1986.

Dutt, Amitava K., "Wage Rigidity and Unemployment: The Simple Diagrammatics of Two Views," *Journal of Post-Keynesian Economics*, Winter 1986–87, 9, 279–90.

Gramlich, Edward M. and Laren, Deborah S., "Migration and Income Redistribution

Responsibilities," *Journal of Human Resources*, Fall 1984, *19*, 489–511.

Hill, Martha S., and Ponza, Michael, "Poverty Across Generations: Is Welfare Dependency a Pathology Passed From One Generation to the Next?," mimeo., Institute for Social Research, University of Michigan, April 1983.

Keynes, John Maynard, *The General Theory of Employment, Interest, and Money*, London: Macmillan 1936.

McLanahan, Sara S., "Family Structure and Dependency: Reproducing the Female-Headed Family" Center for Demography and Ecology Working Paper 85–23, University of Wisconsin-Madison, July 1985.

Murray, Charles, *Losing Ground*, New York: Basic Books, 1984.

______, "White Welfare, White Families, 'White Trash'," *National Review*, March 28, 1986, *38*, 30–34.

Russakoff, David, "A Profound Changing of the Guard Being Felt in Indian Country," *The Washington Post*, January 12, 1983, A2.

Sandefur, Gary D., and McKinnell, Trudy, "American Indian Intermarriage," *Social Service Research*, forthcoming 1987.

Southwick, Lawrence, Jr., "Public Welfare Programs and Recipient Migration," *Growth and Change*, October 1981, *12*, 22–32.

Stewart, James B. and Scott, Joseph, "The Institutional Decimation of Black American Males," *Western Journal of Black Studies*, Summer 1978, *2*, 82–92.

Vedder, Richard and Gallaway, Lowell, "Rich Man, Poor Man: Disincentives in a Transfer Society," paper prepared for Symposium on the Political Economy of the Transfer Society, March, 1986.

Economic Theory and Working Class Poverty Towards a Reformulation

By DAVID H. SWINTON*

Historically, poverty rates for minority individuals and families have been substantially higher than poverty rates for nonminorities. Moreover, after several decades of decline, poverty rates have been increasing for the last decade for both minority and nonminority individuals. The traditional gap between minority and nonminority poverty rates arises primarily within the working class and is largely attributable to differences in minority and nonminority labor market earnings. It also seems clear that much of the recent increase in poverty among the working class is a direct result of increasing employment problems and declining real wage rates. Minority individuals, especially minority males, have been particularly hard hit by these recent labor market trends.

Explanations of poverty and racial differentials in poverty rates among the working class derived from conventional economic theory have correctly emphasized limited earnings in the labor market as the primary determinant of individual poverty and differences in group poverty rates. However, it is the contention of this paper that the conventional explanation of how labor markets generate poverty and poverty rate differentials among the working class is seriously limited and flawed, and thus provides a poor basis for generating good antipoverty policy advice. I propose an alternative view of how labor markets work to generate poverty that provides a richer basis for generating good policy advice.

I. Conventional Explanations for Working Class Poverty

The conventional analysis is based primarily on the interaction of supply and demand forces in a competitive market environment. A simple labor market model assumes the existence of a labor supply function that is based on the optimizing decisions of households, a labor demand function that is based on the optimizing decisions of business firms and a market equilibrium condition. The model may be written as follows:

$$(1) \quad Si = \text{solution to max } u(W, y, z)$$

$$\text{subject to} \quad Si + Zi = Hi; \quad Yi = WSi$$

$$(2) \quad Dj = \text{solution to max } Pj(W, Lj, Kj)$$

$$(3) \quad D(W) = L(W) = S(W),$$

where Si = a labor supply vector for household i; u = a household utility function; W = a vector of prevailing wage rates; Hi = total time available in household i; Zi = leisure time for household i; Yi = total earnings for household i; Dj = labor demand vector for firm j; Pj = a profit function for firm j; Lj = vector of total employment for firm j; Kj = all other factors which influence profits; D = vector of aggregate demand for labor; S = vector of aggregate supply of labor; and L = vector of aggregate employment. All labor vectors are defined over the various types or qualities of labor.

Using this model the conventional analysis can be readily sketched. Given the prevailing wages, each household will choose to supply the quantity of each type of labor that it possesses which will maximize its utility. Aggregating these labor supplies over

*Policy Center-Clark College, Atlanta, GA 30314. Support for the research on which this paper is based was provided by The Ford Foundation. I thank Barbara Jones, Gilbert Moore, and Peter Jackson for helpful comments. I alone bear responsibility for the contents.

all households yields an aggregate supply of labor such that each household is as well off as it can be in the labor market given the quantity and quality of labor it possesses and the prevailing wage structure.

Similarly, each firm will demand the quantity of each type of labor that will maximize its profit given the prevailing wages. The firm will in general hire each type of labor up to the point where they make equal marginal contributions to output per unit of marginal cost. Aggregating these profit-maximizing labor demands across all firms gives the aggregate demand for labor. Again, this aggregate demand is such that no firm can be made any better off at prevailing wages.

The most important and, in my view, the most fatal assumption of the standard model is that market processes operate so as to ensure that the market-clearing condition defined in equation (3) obtains in equilibrium. Equation (3) implies that the market cannot attain equilibrium unless both households and business firms are simultaneously in equilibrium. While orthodox theory is not specific on the labor market process or institutions which bring about the market clearance of equation (3), the process is usually stated as one of wage adjustments.

The disequilibrium case that has some relevance for understanding poverty is when $D(W)=L(W)<S(W)$. This would mean that at least one household was not able to supply all of the labor of some type that it desired to supply. Consequently, any such household would have lower income (LiW) than it would have if it were in equilibrium at (SiW). The lower income would be brought about either by involuntary unemployment or underemployment.

According to the orthodox analysis, if Li were less than Si for some subset of households i, then wages would adjust downwards. This downwards adjustment in wages would lead to a downward adjustment in Si and an upward adjustment in D. These adjustments would continue for every type of labor for which $Li<Si$ until equality of Si and Di obtained. In the orthodox analysis, labor market equilibrium cannot attain until labor markets clear, and all households and firms are satisfied with the situation given the wage rate.

Orthodox analysis also implies that a unique set of wages will emerge from a given market situation. Each wage Wj is the precise wage that clears the market for labor type j. Any increase in the prevailing market-clearing wages will result in unemployment and throw the market out of equilibrium. Thus, the wages received by all types of labor are the best that can be attained given the prevailing conditions.

Using this model, poverty can be explained in a fairly straightforward fashion. First, assume a poverty income standard for each family (Ypi) based on family size and whatever other factors systematically influence need. The poverty standard is presumably set to the income level that enables each family to meet minimum standards of consumption. Ignoring other income sources, each working class family's equilibrium income is given by

$$(4) \qquad Yi=SiW.$$

If $Yi<Ypi$ then the family is poor, and if $Yi>Ypi$ then the family is nonpoor.

The poverty of a working class family is thus strictly determined by its ownership of labor and its willingness to supply that labor, *ceteris paribus*. Any family is poor because it owns or supplies too small a quantity or too low a quality of labor to earn an above-poverty-level income. In the orthodox analysis, no unemployment or underemployment exists in equilibrium. All are assumed able to supply whatever quantities they desire of each quality of labor they own given the prevailing wages.

The orthodox theory implies that the income of any family Yi can differ from the income of any other family Yj if and only if Si does not equal Sj. Thus, differences in the levels of poverty experienced by different families subject to the same poverty standard results strictly from differences in the ownership of or willingness to supply labor. Thus, racial differences in the rates of poverty of families of equal size and composition result strictly from racial differences in their ownership or willingness to supply labor. Aggregate racial differences could also be influenced by racial differences in family size and composition distributions. However, this factor is outside of the scope of this paper and will not be considered further.

Some orthodox theorists entertain the idea of racial discrimination as a partial explanation of racial differences in labor market outcomes. However, since racial discrimination can be shown to be unstable in an orthodox market environment, this explanation is not consistent with the basic orthodox framework. Racial discrimination in the labor market cannot be considered an important cause of racial differences in poverty rates without calling the entire orthodox analysis into question.

The poverty of orthodox theory as a basis for advising policymakers is apparent. The theory leads to the conclusion that there are no labor market policies other than *laissez-faire* necessary to generate the "optimum" level of poverty from labor market operations. There is no unemployment or underemployment that will not be corrected efficiently by market processes. Moreover, since the existing W is the market-clearing wage structure, no other wage structure is possible without increasing unemployment. Orthodox analysis leads to the conclusion that market processes are best left alone. And although tax and transfer policies are possible, they risk distorting labor supply and lowering economic efficiency. Existing poverty levels for working-class families take on the character of being optimum poverty levels.

Only policies that work outside the labor market would be beneficial. Improving the human capital of poor workers could shift the labor supply functions so as to reduce poverty rates. This mechanism would, in general, result in a reduction of the wages of high-wage workers and an increase in the wages of low-wage workers which could reduce poverty rates. Improving the work ethics of the poor could also be recommended. In general this could lead the poor to supply more labor. This would lower the wage rates but could reduce poverty if the gains from increased employment offset the losses from lower wages. Finally, economic development could be recommended. This could increase the demand for labor and thus increase earnings by increasing both wages and employment levels.

However, these types of recommendations must also be viewed with skepticism when they emanate from an orthodox framework because the orthodox assumptions of rationality and competition imply that firms are already making optimum investment decisions, and that households are already acquiring optimum quantities of human capital and supplying optimum amounts of labor. If we presume that these decisions are being made in an irrational fashion, or that there are significant barriers to making these decisions rationally, then this would call into question the entire approach of the orthodox analysis.

II. An Alternative Formulation

The primary problem with orthodox labor market theory is that it is not consistent with empirical reality. In particular, it seems clear that contrary to orthodox predictions, unemployment and underemployment are problems which are not automatically corrected by labor market processes. Substantial unemployment and underemployment exist simultaneously with apparent labor market equilibrium. It also seems empirically clear that substantial differences in poverty levels exist between individual working-class families who apparently are capable and willing to supply equivalent quantities and qualities of labor. It is also empirically apparent that substantial racial differences in poverty rates persist for long periods of time which are not related to either differences in supply behavior or differences in human capital. Thus, it seems apparent that the orthodox theory fail to provide predictions that are consistent with easily observable empirical reality.

As suggested above, the primary flaw in the orthodox model is with the market equilibrium condition. This condition implies that labor markets clear for both households and firms. The alledged market-clearing mechanism is wage adjustments. I suggest that the orthodox wage adjustment process does not exist in most real world labor markets. In particular, I suggest that during a policy relevant time interval, an excess supply of labor has no appreciable impact on wages and thus cannot in general influence the demand for labor. In other words, I suggest that the labor market cannot be analyzed as if the conventional model applies when there is slack demand for labor.

The primary reason for the above conclusion is empirical. Namely, no labor market institutions exist that allow individual workers to bid the market wage down for all workers in order to obtain employment. It is also apparent that such a labor market process has never played a prominent role on a day-to-day basis in U.S. labor markets.

There are, however, several other reasons for rejecting the orthodox market-clearing mechanism. First, most workers are wage takers and generally accept employer wage offers rather than make their own. Second, under normal circumstances, there is no labor market institution which would permit recontracting for employed workers because unemployed workers are willing to work for a lower price. Without the ability to recontract, the savings from negotiating a slightly lower wage with one additional worker would be too small to motivate any but the smallest employers. Third, it is also apparent that most medium-sized firms that pay good wages could replace a large portion of their existing work forces with employees who are willing to work for less at any given time, but firms seldom do this under ordinary circumstances. In fact, it seems clear that most employers are wage makers. They set their wage to the level consistent with their long-term labor market strategy.

Labor demand arises from the profit-maximizing decisions of firms and are based primarily on their sales expectations given the existing wage structure. Unless the sales expectations change, labor demand will not change. In general, therefore, labor demand is primarily determined by developments in the goods market and not by developments among unemployed workers.

When there are excess supplies of labor, employers are already hiring all of the workers they need or desire to carry out their production goals. Therefore, they have no incentive to alter existing employment levels. On the other hand, unemployed workers can seldom do anything effective to alter existing levels of demand. Thus, a situation of excess supply of workers in general or for specific categories must be considered a market equilibrium in the sense that there are no market forces capable of altering such an outcome.

Thus, I suggest that equation (3), the equilibrium condition, should be changed to the following:

$$(3') \qquad D(W) = L(W) \leqq S(W).$$

This condition implies that a labor market equilibrium only requires employers demand for labor to be satisfied. Workers may not be satisfied, but this will in general have no significant impact on the prevailing wage or on the demand for labor. In other words, labor markets do not have to clear in equilibrium.

This new equilibrium condition also implies that W does not emerge from a labor market tantonement process. Thus, W need not be either unique or optimum. We would expect wages to be roughly comparable between firms that compete in the same goods markets and have the same level of efficiency. However, wages may and generally do differ across industries and between more efficient and less efficient firms in the same industry. In general, less efficient firms will be expected to pay lower wages.

The wage structure emerges out of a complex and concrete socio-historical process. Each firm adjusts its behavior to whatever wage structure prevails. However, once a given wage structure gets established either for a firm or an industry, market forces will tend to perpetuate that particular structure. Indeed, I suspect that there is a kind of economic law of inertia that tends to keep an economy flowing along the same tract in the absence of a significant external force. In any case, the implication of this reformulated model is that the wage structure is arbitrary in the sense that the relative wages of different jobs will have no necessary bearing to the relative capabilities of the workers who are employed in the various jobs.

This model is consistent with the empirical finding that, in the long run, workers may adjust to a prolonged shortage for a particular type of labor by altering their supplying behavior. This response is particularly likely for arduous jobs which require expensive training. However, this will not alter the fact that the behavior of the workers will have no immediate impact on the employment levels of firms. It is also possible that firms may lower their starting wages given a condition of surplus labor if they feel that this will not interfere with their overall labor market strategy. However, this possibility does not alter the basic conclusions since these employer wage reactions will not alter their employment plans so long as no change in sales expectations occur.

A simple alternative to the orthodox model can thus be constructed from equations (1), (2), and (3′). Whenever, the strict inequality holds in (3′), there will be unemployment, underemployment, or both. Thus, some individual households will have actual employment vectors (Li) that are less than their available supplies (Si). Some of the labor in such households will be unemployed or underemployed. Moreover, because wages are not necessarily equal for the same work in different firms, an individual i could receive lower wages for the same type of employment than another individual j (i.e., $Wi < Wj$).

It should also be apparent that when the labor market attains an equilibrium with the strict inequality holding, and when there are differences in the wages paid in different firms or industries for the same type of work or work requiring the same labor capabilities, the labor market outcomes cannot be completely determined by the labor market characteristics of individual workers. Thus, workers must compete for jobs and a large number of primarily sociological factors will determine which workers will have the most labor market success. These factors include but are not limited to contacts, credentials, luck, appearance, sex, test-taking ability, and race. Under these conditions, there will be limited labor market incentives to eliminate discrimination.

The reformulated model provides a much richer array of possible explanations for variations in poverty levels. The level of earnings for the ith household will now be determined by its actual employment (Li) and wage rates (Wi) rather than the available supply of labor in the household (Si) and the highest prevailing wage (W). Thus,

$$(4') \quad Y'i = WiLi \leqq WSi; \quad Wi \leqq W; \quad Li \leqq Si.,$$

Poverty may arise in this model for several reasons. First, prevailing wages may be suboptimum in the sense that greater poverty than is required for efficiency is generated by the existing wage vector than some alternative wage vector. Second, some individuals may obtain a wage lower than the maximum available for the labor they supply. Third, there may be unemployment or underemployment. Fourth, some individuals may have a lower supply of labor (quantity or quality) than required for above poverty earnings. This latter explanation is the only explanation that can arise in the orthodox model.

A household i could be poorer than a household j of the same size and composition for several reasons. First, greater poverty for the ith household can be due to lower labor supplies as in the orthodox analysis (i.e., $Si < Sj$). Second, greater poverty for the ith household could be due to greater levels of unemployment or underemployment (i.e., $(Si - Li) > (Sj - Lj)$). Third, greater poverty levels could arise for household i because of receipt of lower wages (i.e., $Wi < Wj$). Unwarranted racial inequality obviously arises whenever blacks are more likely to be unemployed, underemployed, or to receive lower wages than equivalent whites. Warranted racial differences still arise as they did in the orthodox model from racial differences in the characteristics of labor supply.

III. Conclusions

An alternative to the orthodox competitive labor market model can be constructed simply by changing the orthodox equilibrium condition from the requirement that workers and capitalists be satisfied with labor market outcomes to the more realistic notion that the capitalists are the only economic agents that need to be satisfied for market-period equilibrium. While the orthodox model rules out any consistent explanation for poverty except limited supplies of labor, the alternative model offers an array of possibilities including unemployment, underemployment, and a suboptimum wage structure. Moreover, while the orthodox model implies that all differences in poverty rates are warranted by differences in supply characteristics, or behavior, the alternative also permits the existence of unwarranted inequality that arises from the competition of workers for scarce opportunities. Sociological factors such as luck, contacts, and discrimination would all be consistent with equilibrium under the alternative formulation and would be the principal factors explaining the different outcomes for productively equivalent workers.

The alternative formulation offers economists a chance to develop, explore, and test more useful models. This should enable economists to deliver more useful and realistic advice about the best strategies to eliminate working class poverty and racial differences in poverty. However, before this potential can be realized, a more complete model capable of supporting detailed empirical analysis must be developed and estimated. I hope that this paper will stimulate empirical work along the lines suggested by the reformulation.

Audit Classes and Tax Enforcement Policy

By SUZANNE SCOTCHMER*

Compliance with the legislated tax code cannot be assured without audits and penalties on unreported income. Although compliance can most cheaply be assured with a small probability of audit (since audits are costly) and large penalties, law or social convention may prohibit stringent penalties, and the enforcement agency must then decide on an audit strategy as its sole choice variable.[1]

Without audit, the enforcement agency cannot observe true taxable income, but it may be able to observe correlates like profession, age, and gross income, as reported independently by the employer. I shall say that such information defines a taxpayer's audit class. The enforcement agency will find it lucrative to condition the probability of audit on audit class, as well as on reported income, particularly if the audit class is a good signal of income.[2]

Even when the enforcement agency undertakes its best audit strategy, allowable penalties may be so low that taxpayers underreport income in equilibrium. In that case, the effective tax code will differ from the legislated tax code, where the effective tax code reflects *actual* payments, including taxes on reported income and the expected value of fines. The proper definitions of tax equity become murky when taxpayers underreport income, especially when some taxpayers are honest. Should equal consideration be given to honest and dishonest taxpayers, as though underreporting income were a legitimate portfolio choice? Without wishing to dismiss this question as unimportant, this paper assumes that taxpayers are amoral, and is concerned with the effective tax code when taxpayers underreport.

The optimal enforcement literature typically assumes that the object of the enforcement agency is to maximize revenue net of costs. This is a reasonable assumption in this age of budget deficits, since an enforcement agency would be replaced if it left unexploited opportunities to enhance revenue. But enforcement policies designed to maximize net revenue affect the equity properties of the effective tax code. The enforcement agency will audit taxpayers with low income reports within an audit class with higher probability than it audits high-report taxpayers, thus making it less attractive for low-income taxpayers to underreport income, and introducing a regressive bias in the effective tax code within each audit class. To see this, suppose the opposite, that the probability of audit rose with reported income. Then high-income people would have an added incentive to underreport income; namely, that underreporting reduces the probability of audit. The enforcement agency can dissuade high-income taxpayers from underreporting (thus increasing revenue) by

†*Discussants*: Kim Border, California Institute of Technology; James Hines, Princeton University; Joel B. Slemrod, University of Minnesota.

*Graduate School of Public Policy, University of California, Berkeley, CA 94720. I thank Joel Sobel and Isabel Sanchez for useful comments and discussion. This research was supported by NSF grant SES-8610021.

[1] Jennifer Reinganum and Louis Wilde (1986) have studied the equilibrium audit strategy when fine rates are fixed and the enforcement agency can deviate from its announced policy once taxpayers have reported their taxable income. Reinganum-Wilde (1985), Kim Border and Joel Sobel (1986), and Dilip Mookherjee and Ivan Png (1986) have studied the optimal choice of tax code, penalty structure, and audit function when fines cannot exceed income and the enforcement agency can bind itself to an enforcement policy.

[2] Carol Jones (1986) studies optimal enforcement of emission standards when the policy can depend on industry "sectors" that are signals of compliance cost. This is a very similar problem.

making the probability of audit rise as reported income falls.[3]

This seems to defy experience, since audit probability seems to increase with reported income. But the observation that probability of audit rises with reported income pools audit classes. Taxpayers with high reported income might have high probabilities of audit because they are in audit classes that signal high income. Nevertheless, high income reports within that audit class may elicit lower probability of audit than low reports.[4] Evidence for this might be that taking more deductions increases the probability of audit.

Equity properties of the effective tax code will reflect the fact that probabilities of audit depend on audit class as well as on reported income. This may well overturn the regressive bias within audit classes, a complication that has not been studied.

There are two reasons taxpayers with the same income may pay different amounts. First, since people underreport, payments differ according to whether the taxpayer is audited. A taxpayer who is caught underreporting pays tax on the true taxable income, plus a fine based on the discrepancy between true and reported income. Second, taxpayers with the same income may belong to different audit classes. Their expected payments differ both because they typically report different amounts of income and because they are audited and fined with different probabilities. This leads to variance not only in the actual payments, but also in the *expected* payments (where the expectation accounts for whether or not an audit occurs) of taxpayers with the same income.

Thus, enforcement policy affects both vertical and horizontal equity. The vertical "inequity" is that expected payments do not rise with income at the legislated rate. The horizontal inequity is that taxpayers in different audit classes with the same true taxable income have different expected payments.

I present a simple model showing that the regressive bias within audit classes may be dominated by effects across audit classes. I take a linear tax code (constant marginal and average tax rates) and assume that taxpayers with high income are distributed among audit classes that signal high income. The distribution of income signals is the same for every income level, except for a shift of location. Then the average tax payment (expected tax payment, divided by true income) decreases with income *within* each audit class, but when one averages the expected payments of taxpayers with the same income who are in different audit classes, their average tax rates *increase* with true income; that is, the tax code is more progressive than stipulated.

Turning to horizontal equity, it is plausible that when the income signal becomes a better predictor.of true income, the variance in expected tax payments among taxpayers with the same true income might decrease, since they might be treated more uniformly. This is unclear in the case studied below, and the effect on horizontal equity of improving the income signal is an interesting open question.

Each taxpayer has a true taxable income i. His true tax liability, as it would be assessed if he were audited, is ti, where t is a linear tax rate. Although the enforcement agency cannot observe i without audit, it observes a signal y that is correlated with i. In addition, the taxpayer reports an amount of taxable income r. The enforcement agency must decide how often to audit taxpayers of each type; that is, it chooses a probability-of-audit function $p(r, y)$. I assume that the enforcement agency can commit to such a policy, and that the taxpayer knows the policy. A taxpayer who is audited and found underreporting must pay tax on the unreported income plus a penalty at rate f. I assume taxpayers are risk neutral and therefore that they choose their reports r to mini-

[3] See Reinganum and Wilde (1986) for a discussion of this result in the case that the agency cannot precommit, and my 1986 paper for the case that the agency can precommit.

[4] My earlier paper shows that when taxpayers are risk neutral, and when the distributions of true income within audit classes differ by a scale parameter, audit classes with higher average income (but the same mass of taxpayers) will be audited more often than those with lower income.

mize expected payments:

$$\text{(1)}\quad \underset{r}{\text{Minimize}}\ tr+t(1+f)p(r,y)(i-r).$$

I refer to the minimum as $\tau[i, y, p(\cdot)]$, which is achieved by the optimal report $r[i, y, p(\cdot)]$ that solves (1).

Letting the audit probability depend on the audit class is the same as choosing a different probability-of-audit function (a function of reported income) for each audit class. Since the agency can treat distinct audit classes completely separately, it is sufficient to analyze policy for one audit class. If $H(i|y)$ represents the measure of true income i in audit class y, the enforcement agency chooses $p(r, y)$ for each y to maximize

$$\text{(2)}\quad \int\tau[i,y,p(\cdot)]\,dH(i|y) - c\int p[r(\cdot),y]\,dH(i|y),$$

where c is the cost of an audit.

Two features of the optimal audit function are immediate. First, no audit probability will exceed $1/(1+f)$, since unreported income then has expected value zero, and any taxpayer that faces a probability of audit this high will report truthfully. Any higher audit probability would therefore have no additional incentive effect, and would be wasteful, since audits are costly. Second, the audit function $p(r, y)$ (with y fixed) is nonincreasing in r. If $r_1 < r_2$, and $p[r_1] < p[r_2]$, then expected payments (1) are greater with report r_2 than with report r_1, for every income level. Hence r_2 will never be reported, and the increasing portion of the audit function could be replaced with a constant function. Techniques developed by Roger Myerson (1981) and Eric Maskin and John Riley (1984) can be adapted to show that the optimal audit function within an audit class can be characterized as follows:[5]

LEMMA: *The optimal audit function within the audit class has the following form: For some ρ^y, $p(r, y) = 1/(1+f)$ for $r < \rho^y$ and $p(r, y) = 0$ otherwise.*

Knowing this, one can calculate that the optimal cutoff ρ^y satisfies

$$\text{(3)}\quad v\equiv\frac{c}{t(1+f)}=\frac{N^y-H(\rho^y|y)}{h(\rho^y|y)},$$

where N^y is the measure of people in the audit class.

Within each audit class, the effective tax code is regressive: low-income taxpayers report honestly and pay tax on their true income, while taxpayers with income greater than ρ^y pay tax on ρ^y. Since the dishonest high-income taxpayers never get audited, their expected payments are just $t\rho^y$.[6] The average payment schedule of high-income taxpayers is $t\rho^y/i$, which decreases with true income i.

This regressive bias is of limited importance if the most important discriminant for audit probability is the audit class, rather than reported income, as when the audit class is a very good predictor of true income. I model this idea in the simplest possible way. True incomes i and income signals y are jointly distributed in the population according to measure $H(i, y)$. For simplicity, assume incomes in audit class y are distributed uniformly on $[y-a, y+a]$ with density $1/2a$; that is, $h(i, y) = 1/2a$ for $y-a \leq i \leq y+a$. Then the marginal distributions of i and y are uniform with marginal densities one,[7] and thus $h(i|y) = h(y|i) =$

[5] I am grateful to Joel Sobel and Isabel Sanchez for pointing out the relevance of these papers.

[6] Thus, only the honest taxpayers get audited! While the enforcement agency could save costs by refusing to audit, it has an incentive to bind itself *ex ante* to performing the audits. Otherwise the lucrative deterrence of underreporting would vanish. The enforcement agency can always do better by binding itself *ex ante* than by reserving the freedom to deviate from its announced audit policy after tax reports are in, provided taxpayers know which regime is in effect. This is because every policy that is available without precommitment is also available as a precommitment policy.

[7] To ensure that we have a finite mass of taxpayers, assume that the support of $h(i, y)$ is some bounded subset of R^2. Provided that i and y are away from the boundary by at least distance a, the marginal density of i is the integral from $i-a$ to $i+a$ of $\int(1/2a)dy = 1$ and the marginal density of y is the integral from $y-a$ to $y+a$ of $\int(1/2a)di = 1$.

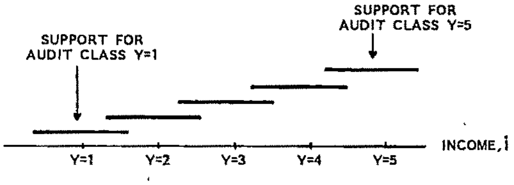


FIGURE 1. TRUE INCOMES AND AUDIT CLASSES

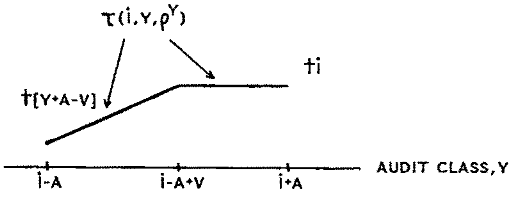


FIGURE 2. EXPECTED PAYMENTS OF TAXPAYERS WITH FIXED INCOME *i*

$h(i, y)$. Figure 1 shows supports for audit classes with increasing average incomes, y. Taxpayers with a particular income i will belong to audit classes with $i - a \leq y \leq i + a$.

The uniform density is inessential. The calculations below regarding vertical equity apply when income signals y are distributed the same for each i except for location; that is, $h(i, y)$ depends only on the difference $y - i$.

The solution to (3) is $\rho^y = y + a - v$. To calculate the effective tax code, we must calculate the average payments (across audit classes) by taxpayers with a given income, i:

$$(4)\quad E_y\tau[i, y, \rho^y] = \int \tau[i, y, \rho^y]\, dH(y|i) = t(i-k),$$

where k is positive.

Here I have substituted ρ^y for the probability-of-audit function in $\tau(\cdot)$ to represent the optimal audit function. $\tau(i, y, \rho^y)$ is graphed in Figure 2. The tax paid is $\tau(i, y, \rho^y) = ti$ if $i \leq y + a - v$, or $i - a + v \leq y$. That is, for large y, income i is a low income in the audit class and the taxpayer will thus report honestly. On the other hand, for small y, $y \leq i - a + v$, income i is a high income in the audit class and since the taxpayer reports the cutoff amount $y + a - v$, $\tau(i, y, \rho^y) = t[y + a - v]$ for such y. The integral can be evaluated by substituting these values for $\tau(i, y, \rho^y)$ in (4), and $(1/2a)$ for $dH(y|i)$ in the domain $i - a \leq y \leq i + a$.[8]

The reason expected payments are less than stipulated tax liability, ti, is simply that taxpayers in low audit classes benefit by being high-income taxpayers in low audit classes and find it advantageous to underreport income. The number of such taxpayers is the same for each income level, and thus the collective amount of unpaid tax is the same, namely tk. The consequence of this fixed underpayment is that the effective tax code is *more* progressive than the stipulated tax code, contrary to what happens *within* audit classes. The average effective tax paid, $t - tk/i$, is increasing with income i.

Turning to horizontal equity, it is clear from Figure 2 that all taxpayers with income i do not make the same payments. One might take as a measure of horizontal inequity the variance in tax payments among taxpayers with the same true income i. One might think that increasing the number of audit classes that contain income i would increase horizontal inequity. Increased variance in y might increase the variance in $\tau(i, y, \rho^y)$ for each i.

Variance in the distribution of y, conditional on i, increases with a, since, conditionally on i, y is uniformly distributed on $[i - a, i + a]$. The end points in Figure 2 shift out. At the same time, the cutoff value of y separating taxpayers that report honestly from those that lie, namely $i - a + v$, shifts left. The result of increasing variance is that more type i taxpayers report honestly! For some parameter values a, the variance in tax paid by taxpayers with income i will actually decrease when they are distributed

[8] If $a < v/2$, then no taxpayers will be audited. Rather, all taxpayers in audit class y will pay tax on the minimum income in the class, $y - a$.

among more audit classes, y.

Many issues about equity arise, which this paper has not addressed. When profession is important in defining the audit class, it is unclear that we really want to define equity according to how much tax is paid on a year-by-year basis. Low income in one year may well be an aberration. To the extent that profession is a stable indicator of *lifetime* earnings, or perhaps of "ability," it would be reasonable to base not only the probability of audit, but also the stipulated tax liability, on profession and not just on income.

During a taxpayer's career, he or she may move from being a low-income taxpayer within the audit class to a high-income taxpayer within the same audit class. The horizontal "inequity" would then vanish over a lifetime.

If there were only one audit class, as when the enforcement agency can observe nothing more about the taxpayer than reported income, then the legislature might be able to undo the regressive bias introduced by enforcement by building more progressivity into the stipulated tax code than is really desired. Such an approach will not work when the probability of audit can depend on the audit class, but the stipulated tax code cannot.

REFERENCES

Border, Kim and Sobel, Joel, "Samurai Accountant: A Theory of Auditing and Plunder," Working Paper, California Institute of Technology, 1986.

Jones, Carol, "Models of Regulatory Enforcement and Compliance," mimeo., University of Michigan, 1986.

Maskin, Eric and Riley, John, "Monopoly with Incomplete Information," *Rand Journal of Economics*, Summer 1984, *15*, 171–96.

Mookherjee, Dilip and Png, Ivan, "Optimal Auditing, Insurance and Redistribution," manuscript, Graduate School of Management, UCLA, 1986.

Myerson, Roger, "Optimal Auction Design," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

Reinganum, Jennifer and Wilde, Louis, "Income Tax Compliance in a Principal-Agent Framework," *Journal of Public Economics*, January 1985, *26*, 1–8.

______ and ______, "Equilibrium Verification and Reporting Policies in a Model of Tax Compliance," *International Economic Review*, October 1986, *27*, 739–60.

Scotchmer, Suzanne, "Equity in Tax Enforcement," Harvard Institute of Economic Research Working Paper 1233, May 1986.

Tax Evasion and Capital Gains Taxation

By JAMES M. POTERBA*

The Internal Revenue Service estimates that in 1985 tax evasion reduced personal income tax receipts by $84 billion, or nearly 20 percent. Unpaid income taxes were 40 percent as large as the federal deficit. The responsiveness of tax compliance to changes in marginal tax rates has attracted significant policy interest in the last two years, since the Tax Reform Act of 1986 lowers marginal tax rates for more than half of the taxpaying population. Three recent studies have used micro data for the United States to investigate the relationship between marginal tax rates and tax evasion. Two of these studies, Charles Clotfelter (1983) and Craig Alexander and Jonathan Feinstein (1986), find sizable marginal tax rate effects. A third study, by Joel Slemrod (1985), finds no effect. This paper provides new evidence on how marginal tax rates affect compliance levels by analyzing the time-series movements in voluntary reporting rates for one type of income, capital gains, between 1965 and 1982.

Two factors make capital gains evasion during this time period a natural experiment in tax compliance. First, the top marginal tax rate on long-term capital gains varied from 20 to 35 percent. Second, capital gains transactions were not subject to information reporting requirements so the potential for evasion was much higher, and the probability of detection much lower, than for other income sources such as wages. This is reflected in higher voluntary reporting rates for wage and interest income, 94.9 and 88.1 percent, respectively, than for capital gains income. The compliance rate for capital gains was only 64.3 percent prior to the recent changes in information reporting rules.

The paper is divided into three sections. The first discusses several methodological issues that arise in interpreting cross-sectional studies of how marginal tax rates affect individual tax compliance and household behavior more generally. Section II presents new empirical evidence on how tax rates affect compliance based on time-series analysis of capital gains reporting rates. The conclusion evaluates the debate over whether the capital gains tax reduction of 1978 was self-financing in light of these findings on evasion behavior.

I. Cross-Section vs. Time-Series Data in Empirical Public Finance

Most previous studies of individual tax evasion, like studies of charitable giving and the realization of capital gains, analyze cross-section data on individual tax returns. Two problems arise in using these data to assess how marginal tax rates affect household behavior. First, it is very difficult to separate income effects from marginal tax rate effects. Most of the dispersion in marginal tax rates is generated by variation in income, so estimated tax rate coefficients may reflect nonlinear income effects rather than tax rate effects. Slemrod is unable to separate income and tax effects with any confidence, and even when estimated tax coefficients are statistically significant, they may not describe the behavioral response to a tax reform. Daniel Feenberg (1982) suggests a potential remedy for this problem and uses interstate variation in income tax rates to identify the impact of marginal tax rates on charitable giving. Additional progress could be made with panel data spanning multiple tax regimes, but such information is not available in the tax evasion context.

*Department of Economics, MIT, Cambridge MA, 02139. I am grateful to Dennis Cox, Berdj Kenadjian, and Thomas Thompson of the IRS Compliance Analysis Group for providing aggregate compliance data, and to Michael Graetz, James Hines, and Lawrence Summers for helpful discussions. This research was supported by the National Science Foundation and a Batterymarch Financial Fellowship. It is part of the NBER Taxation Program.

The second problem with cross-sectional data is that much of the variation in marginal tax rates *conditional* upon income results from household choices. These choices may be correlated with omitted individual characteristics that also affect the behavior, such as evasion, under investigation. For example, married taxpayers face lower marginal tax rates than single taxpayers with identical earnings, but marriage may affect a taxpayer's compliance behavior through channels other than the marginal tax rate. Unobserved characteristics that affect a taxpayer's level of charitable giving or his demand for home ownership will also affect marginal tax rates, and they may affect the proclivity to evade taxes as well. Because tax returns contain minimal demographic data, controlling for these omitted characteristics is extremely difficult.

Relying on time-series analyses of tax rates and household behavior is appealing precisely because the experimental variation derives from changes in the tax code. Time-series studies encounter other difficulties, however. First, it is difficult to summarize the tax system in one or a few variables. For most behavioral decisions, there is enormous heterogeneity in the marginal tax rates facing different taxpayers. For capital gains, marginal rates vary both because of differences in investors' noncapital gains income that affect their marginal tax rates on capital gains, as well as from particular gain and loss realization patterns (see my forthcoming paper). Fortunately, the capital gains tax reforms of the last two decades affected the marginal tax rates facing most investors in similar ways, so they may still be useful for tax research.

The second problem with time-series data is the inevitable difficulty of controlling for other factors that affect taxpayer behavior. Two factors are particularly significant for analyzing capital gains tax evasion: (*i*) Intertemporal variation in tax enforcement is potentially very important. In 1965, the first year of my sample, 4.6 percent of individual tax returns were examined by IRS revenue agents and auditors. By 1982, the last year, only 1.5 percent of returns were examined. There have also been changes in tax shelter enforcement that may affect capital gains reporting. In 1973, the IRS began an enforcement program directed at oil and gas shelters; the program was expanded to other shelters in the late 1970's. (*ii*) The composition of gains also affects capital gains compliance. Some types of gains, notably those on stocks and bonds, have voluntary reporting rates of nearly 90 percent. Corporate stock accounts for only a third of capital gains, however, and many other transactions such as sales of real property have much lower compliance rates (see Thomas Thompson, 1987). The mix of gains has shifted through time, with real estate transactions becoming increasingly important, and this may affect the compliance level.

Despite these difficulties, time-series data provide a new source of evidence on taxpayer behavior. At worst, they constitute a useful validation of the estimates from cross-sectional studies, and at best they may yield more reliable estimates of how structural tax reform will affect household behavior. The next section analyzes time-series data on the capital gains tax voluntary reporting percentage (VRP) to explore how marginal tax rates affect tax evasion.

II. Time-Series Evidence on Tax Rates and Tax Compliance

The IRS estimates the fraction of realized capital gains that are reported on tax returns (*VRP*) as part of each Taxpayer Compliance Monitoring Program (TCMP) survey. There were six TCMP surveys between 1965 and 1982, and the estimated *VRP*s varied from a high of 83.2 percent in 1965 to a low of 61.1 percent in 1976. These data, plotted as the dashed line in Figure 1, are described in more detail in Internal Revenue Service (1983).

I investigate the relationship between the capital gains *VRP* and two measures of the marginal tax burden on capital gain realizations. The first, *MTR*1, is the maximum statutory tax rate on long-term gains. I ignore a variety of complicated capital gains tax provisions that affected a very small fraction of investors during the mid-1970's (see Lawrence Lindsey, 1987, for a more detailed

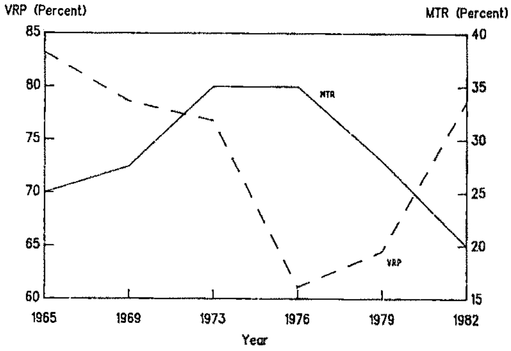


FIGURE 1. MARGINAL TAX RATES AND VOLUNTARY REPORTING PERCENTAGES, 1965–82

discussion of provisions involving the alternative minimum tax and the maximum tax on earned income). This tax rate series is plotted as the solid line in Figure 1. The second tax rate series, *MTR*2, is a weighted average of actual marginal rates on realized long-term gains computed by Lindsey (1987). Its movements are similar in direction, but less dramatic, than those in *MTR*1. Although Joseph Stiglitz (1983) and George Constantinides (1983) emphasize the impossibility of distilling the capital gains tax system into a single marginal tax rate that affects household behavior, I argue elsewhere (1987) that the marginal tax rates on long-term gains realized by a majority of investors move in tandem with these series. Some investors may develop trading strategies that shelter gains and therefore face zero marginal tax rates on capital gains, so the tax reforms have no effect on them. Very few, if any, investors faced reduced capital gains tax rates as a result of the legislation that raised the top marginal rate.

I estimate regression equations linking the logarithm of the voluntary reporting percentage with the log of the marginal tax rate and a time trend, the latter included to capture changes in enforcement, tax compliance mores, and other factors. The results for the two basic equations are shown below, with standard errors in parentheses:

$$(1)\quad \ln(VRP) = \underset{(0.250)}{-0.680} - \underset{(.197)}{.410}*\ln(MTR1) - \underset{(.022)}{.044}*TIME \qquad R^2 = .68,$$

$$(2)\quad \ln(VRP) = \underset{(1.065)}{-1.787} - \underset{(.652)}{.979}*\ln(MTR2) - \underset{(.031)}{.042}*TIME \qquad R^2 = .39.$$

Equation (2) is estimated by instrumental variables, since *MTR*2 is based on reported capital gains, with the maximum statutory rate (*MTR*1) as an instrument.

Both equations suggest important marginal tax rate effects on the tax evasion decision. The first equation implies that a 1 percent change in the marginal tax rate raises the reported tax base by .4 percent. A change like the 1978 tax reform, which lowered the marginal tax rate from 35 to 28 percent, would therefore raise reported capital gains by roughly 9 percent. The second equation suggests an even larger tax rate effect, with reported gains displaying a unit elasticity with respect to the marginal tax rate. The 1978 tax reform reduced *MTR*2 from 21.8 to 17.7 percent, so the second equation predicts a 20.3 percent change in the reported tax base.

Although these equations are estimated using only six observations, the *MTR* coefficient is statistically significant at the .15 level in the first equation and at just below the .20 level in the second. The trend variable in these equations may be capturing changes over time in tax enforcement. To allow for this possibility I replaced the trend variable with the fraction of individual tax returns that were audited. The estimated marginal tax rate coefficient changes very little between (1) and this specification, but the coefficient on enforcement probability is statistically insignificant. Its point estimate is large, however, and suggests that a 1 percentage point increase in the examination probability raises tax compliance by about 4 percent.

To control for the possibility that attitudes toward tax evasion in general had evolved through time in ways that were spuriously correlated with the capital gains tax rate, I also estimated an equation controlling for the level of tax evasion on other types of capital income:

$$(3)\quad \ln(VRP_{cg}/VRP_{int\&div}) = \underset{(.108)}{-.250} \; \underset{(.262)}{-\;.368}*\ln(MTR_{cg}/MTR_{div}) \; \underset{(.029)}{-\;.046}*TIME \quad R^2=.51.$$

The dependent variable is the log of the *VRP* on capital gains, divided by the *VRP* for interest and dividend income. The tax variable is *MTR*1 divided by a weighted average marginal tax rate on individual dividend income, calculated by Martin Feldstein and Joosung Jung (1987). This equation tests the hypothesis that changes in the relative tax burdens on different types of capital income, capital gains vs. interest and dividends, lead to changes in the relative compliance rates on the different income sources. The results support this view. Although the standard error on the tax rate variable is now somewhat higher than in equation (1), the coefficient changes relatively little and the implied elasticity of the tax base with respect to the marginal tax rate is .37.

The time-series evidence on the sensitivity of tax evasion to marginal tax rates is surprisingly similar to the findings of cross-sectional studies. Evaluated at the 1976 values of the compliance level and marginal tax rate, the estimates in equations (1) and (2) imply elasticities of unreported capital gains with respect to marginal tax rates of .64 and 1.54, respectively. By comparison, Clotfelter reports an elasticity of unreported taxable income of 1.46 with respect to the marginal tax rate for high income filers, the group most comparable to the taxpayers reporting capital gains. The comparison with Alexander and Feinstein is more difficult, because they report primarily probit results on the discrete choice of whether or not to evade. For a taxpayer with total taxable income of \$100,000 and \$20,000 of taxable capital gains, their estimates imply that reducing the taxpayer's marginal tax rate from .45 to .33 (as the Tax Reform Act of 1986 does) would reduce the probability of tax evasion from .72 to .55. If all individuals who evade fail to report the same amount of income, this would imply an elasticity of unreported income with respect to marginal tax rates of 1.53.

The estimates from (1) and (2) are only suggestive for two reasons. First, they fail to control for changes through time in the composition of capital gains, principally the increased importance of residential capital gains. If anything, this would induce a downward trend in the measured *VRP* over time, making the 1982 compliance increase even more difficult to explain. Second, the estimates make only a crude correction for varying enforcement patterns. The results do however support earlier studies that find a significant role for marginal tax rates in determining tax compliance.

III. Tax Policy Implications

The influence of marginal tax rates on taxpayer compliance is a central issue in assessing the revenue effects of tax reform. Total revenue raised from a tax is the product of the tax rate and the reported tax base:

$$(4)\qquad T=\tau*\nu(\tau)*B(\tau),$$

where the reported base is the product of the true tax base, $B(\tau)$, and the voluntary reporting percentage $\nu(\tau)$. The revenue effect of a tax reform can therefore be decomposed into three parts, a rate effect, a reporting effect, and a behavioral effect on the true tax base. Although the elasticity of the reported capital gains base $\nu(\tau)*B(\tau)$ with respect to tax rates has been a subject of substantial debate, there has been virtually no discussion of the effect of tax rates on tax compliance $\nu(\tau)$ as opposed to gain realization, $B(\tau)$. Most previous studies treated $\nu(\tau)$ as fixed at one in interpreting their findings on the distortions due to the capital gains tax. The time-series estimates using *MTR*2 imply that a 1 percent change in the marginal tax rate leads to a 1 percent change in reported income, so even without any change in the true tax base, $B(\tau)$, capital gains tax cuts would be essentially self-financing. The estimates using *MTR*1 imply that only half of the revenue lost through reduced rates is made up by increased reporting.

The estimates from the last section can be used to evaluate the relative importance of reporting effects, $\partial \ln \nu / \partial \ln \tau$, and realization effects, $\partial \ln B / \partial \ln \tau$, in accounting for the elasticity of the capital gains tax base with respect to tax rates. First, consider the cross-sectional data on the elasticity of reported realizations. Studies using panel data to disentangle temporary and permanent changes in marginal tax rates suggest an elasticity of realized long-term gains with respect to marginal tax rates of between -1.2 and -2.2 (see U.S. Treasury, 1985, and Gerald Auten and Clotfelter, 1982). The compliance effects in the last section imply that between one-quarter and one-half of these effects could be due to variability in taxpayer reporting, not to changing realization behavior. Feldstein et al. (1980) estimated larger realization elasticities than those from the panel data studies so reporting effects explain a smaller fraction of their results.

There have also been time-series studies of how capital gain realizations respond to marginal tax rates. Lindsey (1987) concludes that a 1 percentage point reduction in the marginal tax rate on capital gains, measured as *MTR*2, will raise realized long-term gains by 5 percent. This implies that a tax change like the 1978 reform would increase the reported capital gains tax base by 20 percent. Evasion effects could account for more than 40 percent of the total effect. Both the micro and time-series estimates suggest that previous studies overstate the behavioral distortions from the capital gains tax.

The problem of distinguishing reporting effects from more substantive behavioral effects arises in other microeconometric tax research as well. Studies of charitable giving that estimate how marginal tax rates affect contribution levels may be capturing in part a tax compliance effect. Richard Fratanduono (1986) reports estimates from the 1982 TCMP showing that charitable contributions were overstated by 10 percent, in comparison with 14.7 percent overstatement in the late 1960's when marginal tax rates were typically much higher. An important but unresolved issue concerns the extent to which the sizable increase in reported taxable income following the marginal rate reductions of 1981 (see Lindsey, 1985) can be attributed to changing compliance patterns.

Recognizing the possibility of capital gains tax evasion, and changing evasion opportunities over time, can also affect analyses of tax distortions. There are wide inter-asset disparities in noncompliance rates. Compliance is much lower for sales of real assets such as business property and personal residences than on corporate stock and bonds (see Thompson). This effectively reduces the tax burden on structures (see Roger Gordon et al. 1987) and also implies that recent initiatives to increase tax compliance by requiring information reporting on most asset sales will alter the relative tax burdens on different assets. Complete analysis of this problem requires integrating work on the deadweight burden due to income tax evasion with work on interasset and intertemporal distortions.

REFERENCES

Alexander, Craig and Feinstein, Jonathan, "A Microeconometric Analysis of Income Tax Evasion," mimeo., MIT, November 1986.

Auten, Gerard and Clotfelter, Charles, "Permanent versus Transitory Tax Effects and the Realization of Capital Gains," *Quarterly Journal of Economics*, November 1982, *97*, 613–632.

Clotfelter, Charles T., "Tax Evasion and Tax Rates: An Analysis of Individual Returns," *Review of Economic and Statistics*, August 1983, *65*, 363–73.

Constantinides, George M., "Capital Market Equilibrium with Personal Tax," *Econometrica*, May 1983, *51*, 611–36.

Feenberg, Daniel, "Identification in Tax-Price Regression Models: The Case of Charitable Giving," NBER Working Paper No. 988, September 1982.

Feldstein, Martin S. and Jung, Joosung, "The Effects of Tax Rules on Nonresidential Fixed Investment," in M. Feldstein, ed., *The Effects of Taxation on Capital Formation*, Chicago: University of Chicago Press, 1987.

______, **Slemrod, Joel and Yitzhaki, Shlomo,** "The Effects of Taxation on the Selling of Corporate Stock and the Realization of Capital Gains," *Quarterly Journal of Economics*, June 1980, *94*, 777–791.

Fratanduono, Richard J., "Trends in Voluntary Compliance of Taxpayers Filing Individual Tax Returns," in *Trend Analysis and Related Statistics: 1986 Update*, Washington: U.S. Department of the Treasury, IRS, March 1986.

Gordon, Roger H., Hines, James R. and Summers, Lawrence H., "Notes on Taxation of Structure," in M. Feldstein, ed., *The Effects of Taxation on Capital Formation*, Chicago: University of Chicago Press, 1987.

Lindsey, Lawrence B., "Taxpayer Behavior and the Distribution of the 1982 Tax Cut," NBER Working Paper No. 1760, October 1985.

______, "Capital Gains: Rates, Realizations, and Revenues," in M. Feldstein, ed., *The Effects of Taxation on Capital Formation*, Chicago: University of Chicago Press, 1987.

Poterba, James M., "How Burdensome are Capital Gains Taxes?," *Journal of Public Economics*, forthcoming 1987.

Slemrod, Joel, "An Empirical Test for Tax Evasion," *Review of Economics and Statistics*, May 1985, *57*, 232–38.

Stiglitz, Joseph, "Some Aspects of the Taxation of Capital Gains," *Journal of Public Economics*, July 1983, *21*, 257–94.

Thompson, Thomas, "1979 Individual Income Tax Capital Gains Income Reporting Noncompliance," in *Trend Analyses and Related Statistics: 1987 Update*, Washington: U.S. Department of the Treasury, IRS, forthcoming 1987.

Internal Revenue Service, *Income Tax Compliance Research, Estimates for 1973–1981*, Washington: U.S. Department of the Treasury, July 1983.

U.S. Department of the Treasury, *Capital Gains Tax Reductions of 1978*, Washington: Office of Tax Analysis, 1985.

Are We a Nation of Tax Cheaters? New Econometric Evidence on Tax Compliance

By JEFFREY A. DUBIN, MICHAEL J. GRAETZ, AND LOUIS L. WILDE*

In 1982, then Commissioner of Internal Revenue Roscoe Egger reported to Congress that legal sector noncompliance with the Federal Income Tax statutes generated an "income tax gap" of $81 billion in 1981, up from $29 billion in 1973. He further projected a gap of $120 billion for 1985 (U.S. Congress, 1982). Perceptions of accelerating noncompliance inspired a crisis mentality within the Internal Revenue Service, Congress, and the tax bar.

The IRS responded in part by funding a major independent study of tax noncompliance via the National Academy of Sciences, and the American Bar Foundation initiated an investigation of its own in 1984. Congress enacted compliance legislation in 1981, 1982, and 1984, and completely overhauled the federal income tax laws in 1986. These enactments added a wide variety of new penalties for noncompliance and strengthened others, dramatically expanded requirements for third-party reporting of information to the IRS, added to the IRS's arsenal of procedural weapons, and adopted everyone's favorite vehicle to combat noncompliance—lower tax rates.

All this clamor and action has taken place in the absence of any solid factual foundation (Graetz and Wilde, 1985). We are not at all certain of the actual decline in tax compliance during the past decade, and even if noncompliance has increased significantly, its causes, and thus appropriate remedies, simply are not known. For example, recent unpublished IRS estimates have significantly reduced Commissioner Egger's projections for 1985—to $92 billion; in fact, the real income tax gap for individual returns is now thought to have fallen from $39.1 billion in 1981 to $36.8 billion in 1986, measured in 1972 dollars. These figures do not support the widespread claims that the American public is becoming a nation of tax cheaters, or that the integrity of the tax system is seriously at risk, but the complete story is much more complex. Not only must there be additional efforts to determine what circumstances imply increased noncompliance, but the effects of recent tax law and penalty changes as well as changes in IRS budgets and audit capacity must also be taken into account. Ultimately this is an empirical story, but valid empirical work must be based on the proper theoretical foundation.

The theoretical basis for the economic approach to tax compliance has, at least until recently, been inadequate, and the limited empirical work based on it is seriously flawed. In this paper we briefly review both, as well as new theoretical and, especially, empirical work on the tax compliance problem.

I. The Economic Theory of Tax Compliance

The contemporary revival of the economic analysis of crime began with Gary Becker's classic 1968 article. While Becker mentioned tax evasion as a potential application of his general model, Michael Allingham and Agnar Sandmo (1972) published the first formal analysis. In their model, the taxpayer's actual income is exogenously given and known only to the taxpayer. A constant proportional tax is applied to reported income, with such reported amounts chosen by the taxpayer to maximize expected utility of net wealth. With some exogenous and constant probability, the taxpayer is "audited." If the taxpayer is discovered to be underre-

*Assistant Professor of Economics, California Institute of Technology, Pasadena, CA 91125; Professor of Law, Yale University, New Haven, CT 06520; and Professor of Economics, California Institute of Technology, respectively.

porting income, a penalty proportional to the amount of undeclared income must be paid in addition to the proportional tax rate.

The bulk of the remainder of the theoretical economics literature on tax compliance consists largely of extensions and refinements of Allingham and Sandmo's model. While many ambiguous results are produced by these analyses, one prediction is universal: an exogenous increase in the probability of detection and conviction or in the penalty rate will increase compliance.

More recent theoretical innovations have attempted to move out of the decision-theoretic framework characteristic of the early tax compliance literature. Of particular interest here are the principal-agent models of Kim Border and Joel Sobel (1987) and Jennifer Reinganum and Wilde (1985) and the game-theoretic model of Graetz et al. (1986). In both of these approaches the IRS is allowed to act strategically, conditioning its audit rules on the information it receives from taxpayers.

Whether the IRS should be included as a strategic actor in theoretical models of tax compliance is of more than technical interest. In assessing empirically the deterrent effects of audits, it is crucial whether the IRS audit selection process turns on taxpayer compliance behavior. If it does, then any empirical model meant to explain taxpayer compliance behavior that treats audit rates as exogenous may be seriously misspecified. In fact, any deterrent effect of audits may be outweighed by a (presumed) countervailing incentive of the IRS to audit most heavily those returns for which expected compliance is the lowest (in light of information received) and thereby produce an observed negative relationship between audits and compliance.

II. Existing Empirical Work

To date, there has been a surprisingly small amount of empirical work on the determinants of tax compliance. Not counting survey work or work attempting to measure aggregate noncompliance, we have found only four relevant studies; Charles Clotfelter (1983), Ann Witte and Diane Woodbury (1984, 1985), Joel Slemrod (1985), and Dubin and Wilde (1986).

Clotfelter analyzed a data set collected originally as part of the 1969 IRS Taxpayer Compliance Measurement Program (TCMP). The TCMP involves detailed "line-by-line audits" of a stratified random sample of taxpayers, which result in income tax assessments regarded by the IRS as "correct." The IRS uses TCMP audits in developing a scoring mechanism (the "Discriminant Index Function," or "DIF") to establish and refine the audit selection decisions it applies to the larger population of taxpayers. The TCMP is a far better technique for learning about the effectiveness of IRS audits than about aggregate noncompliance (Graetz and Wilde), but it is nevertheless one of the best sources of data currently available for estimating noncompliance.

Using raw TCMP data, Clotfelter investigated the relationship between marginal tax rates and tax evasion for three classes of taxpayers (nonbusiness, nonfarm business, and farm). For each group, he regressed the log of underreported income on a measure of the effective marginal tax rate, after-tax income, wages as a proportion of adjusted gross income, interest, and dividends as a proportion of adjusted gross income, and several socio-demographic variables. The average audit rate for each taxpayer class was not included as an independent variable since, as Clotfelter put it, "the probability [of audit] for any tax return in a given class is a function of its reported items" (p. 336); in other words, there is a potential simultaneity problem that makes it inappropriate to use audit rates as exogenous explanatory variables in an equation meant to explain compliance with the tax laws.

Clotfelter found that both the level of after-tax income and marginal tax rates have significant negative effects on compliance. While these results are interesting, they should be used with caution. Clotfelter tried to avoid the simultaneity issue by leaving audit rates out of his model, but his model is still misspecified if audit rates affect compliance. In any event, since he left audit rates out of his analysis, Clotfelter's work implies nothing about their deterrent effects.

Witte and Woodbury (1985) explicitly attempt to analyze the effects of audit rates and sanction levels on compliance using a data set provided to them by the IRS. This data includes a percentage compliance variable related to 1969 returns filed in 1970 (estimated by the IRS from DIF scores, not actual IRS audits), IRS agency variables such as audit rates and sanction levels, and a host of demographic and socioeconomic variables, all aggregated to the three-digit zip code level. Separate equations were estimated for each of seven audit classes, defined by income level (low, medium, or high) and by type of return (1040 only, Schedule C or F present, Schedule C and F not present), using seemingly unrelated regression. In particular, the estimated 1969 percentage compliance variable was regressed on a constant term and 36 explanatory variables, including audit rates for 1967, 1968, and 1969 within the audit class, and for all other audit classes.

A detailed discussion of Witte and Woodbury's work can be found in Dubin and Wilde. Two major problems with it are the numerical properties of their data set are unsatisfactory, and many of the agency variables are likely to be endogenous so that their model is misspecified.

These problems perhaps explain some of the peculiar results obtained by Witte and Woodbury. In their 1985 paper, for example, they report selected results for 3 of the 7 audit classes. For these audit classes, reported mean elasticities of percentage compliance with respect to "audit rates" range from .002 to .02, approximately. However, by referring to their 1984 working paper, one finds first that these elasticities are obtained by summing the coefficients, when significant, on all 6 of the audit variables (1967, 1968, and 1969 audit rates within each audit class and for all other audit classes). Second, only one of the 1969 within-class audit rate variables is significant and it has a negative sign, 6 of the 7 1968 within-class audit rate variables are significant and half have a negative sign, and 6 of the 7 1967 within-class audit rate variables are significant, but all have a positive sign. It is difficult indeed to conclude from these results that increases in audit rates increase compliance.

Slemrod takes a different approach in his analysis of tax avoidance. He notes that tax liability is a step function of taxable income for most taxpayers, the step-size being $50 in 1977. He shows that noncompliers, theoretically, have an incentive to report income levels near the upper end of the relevant step range. Using 1977 TCMP data, Slemrod regresses the taxpayer's position within the relevant $50 bracket (a number from 1 to 50) on several factors. A tendency to be located higher in the relevant $50 bracket is shown to be positively associated with higher marginal tax rates, being less than 65-years-old, being married, and the presence of certain "fungible items." But Slemrod's approach is of limited value at best; it cannot get at the degree of tax evasion even if it is present, as he hypothesizes.

The most recent empirical study of tax noncompliance using microeconomic data is by Dubin and Wilde. They analyze a subset of the 1969 data set used by Witte and Woodbury, augmented with data taken from the 1969 *Annual Report of the Commissioner of Internal Revenue*. The dependent variable is the same as that used by Witte and Woodbury, an IRS estimate (based on DIF scores) of the percentage compliance rates for individuals in each of the 7 audit classes described above.

The explanatory variables are the 1969 within-class audit rate; 3 variables that have been thought to reflect opportunities to evade: the unemployment rate, the percentage employed in manufacturing, and the present self-employed; and 3 variables that the literature (principally surveys) suggests are important: the percentage of the population over 65-years-old, the percentage of persons over 25 with at least four years of high school education, and the percentage of nonwhite population. Following recent theoretical work (and empirical work elsewhere in the economics of crime literature), Dubin and Wilde treat the audit rate as being potentially endogenous. This hypothesis is tested using an instrumental variables procedure, the "instruments" being the number of criminal fraud investigations initiated in 1970 per 1968 return filed in 1969, the percentage of taxpayers receiving a first or second notice in 1969 indicating that taxes were due, and the IRS budget per tax return filed. Dubin and Wilde regard the last of these as a good instrument but discuss possible shortcomings of the first two.

In 4 of the 7 audit classes (low-income nonbusiness with a standard deduction, low-income business, and both high-income classes), the audit rate was found to be endogenous. In all audit classes a deterrent effect of audits on noncompliance was found, but in 3 of the 4 cases in which audits were found to be endogenous, the deterrent effect was dominated by the countervailing incentives for the IRS to audit most heavily those returns with the greatest expected noncompliance, so that, in equilibrium, audit rates were negatively related to compliance for these audit classes.

TABLE 1—COMPLIANCE, COLLECTIONS, AND AUDIT EQUATIONS, 1978–85

Independent Variable[a]	Dependent Variable		
	PBANG	*ICR*[c]	*IAR*(−1)
ONE	−12.945	−5.142	10.765
(1.00)	(−4.688)	(−3.809)	(3.595)
IAR(−1)	−0.008	0.003	–
(1.65)	(−1.449)	(1.000)	
PIAR[b]	0.881	0.181	–
(1.741)	(5.403)	(2.274)	
PERED(−1)	3.607	−0.167	−0.813
(0.68)	(4.224)	(−0.401)	(−0.972)
PEROLD(−1)	−0.940	1.675	−2.293
(0.42)	(−0.659)	(2.401)	(−1.759)
UR(−1)	−6.148	2.811	0.908
(0.073)	(−2.645)	(2.473)	(0.417)
PICAP(−1)[c]	−1.162	1.548	0.009
(5.33)	(−5.367)	(14.614)	(0.458)
PICAP2(−1)	0.006	−0.007	−0.001
(29.73)	(4.431)	(−10.963)	(−0.733)
PMAN(−1)	−1.145	3.049	−1.460
(0.19)	(−1.409)	(7.672)	(−2.154)
TIME	0.187	−0.001	−0.008
(81.5)	(6.301)	(−0.367)	(−2.961)
BPR(−1)[c]	–	–	289.965
(0.00042)			(6.745)
PIRF(−1)	–	–	−3.267
(0.587)			(−2.927)
Number of Observations	400	400	400
R-squared	0.298	0.632	0.247
Mean of Dependent Var.	0.684	2.234	1.741

[a] Mean values are shown in parentheses below variable names; *t*-statistics (below coefficients) while qualitatively similar to instrumental variable estimates, are not identical.

[b] *PIAR* is the predicted value of *IAR*(−1) from the audit equation.

[c] Measured in thousands of 1972 dollars.

III. New Empirical Work

The *Annual Report of the Commissioner of Internal Revenue* contains a wealth of data that has not to our knowledge been exploited by researchers. The typical report gives state-level information for each type of tax (individual and corporate income, estate and gift, etc.) regarding total collections, number and amount of refunds, number of returns filed, number of returns examined, additional tax and penalties recommended after examination, and costs incurred by the IRS. We have assembled data from these *Annual Reports* for 1977 through 1985, and we expect eventually to add additional years. We have also obtained for these years data on socioeconomic and demographic variables similar to those used by Dubin and Wilde.

We have just begun to explore this rich data set, but are able to report here three preliminary results. Following Dubin and Wilde, we use the percentage return per audit for individual returns (additional tax and penalties from audits ÷ total collections, per million audits, in 1972 dollars) (*PBANG*) as the dependent variable. Our independent variables are lagged values of the audit rate (examinations per 100 returns filed) (*IAR*); percent of the adult population with a high school education (*PERED*); percent of the population over 45 (*PEROLD*); per capita income (*PICAP*) and its square (*PICAP2*); the unemployment rate (*UR*); percent of the work force employed in manufacturing (*PMAN*); and a time trend (*TIME*). We allow for endogeneity of the audit rate using the budget per return (*BPR*) and the percent of individual returns filed (*PIRF*) as instruments. The time-series results are broadly consistent with Dubin and Wilde's cross-section results: 1) the audit rate is endogenous as indicated by the significant coefficient of the predicted audit rate (*PIAR*) in the compliance equation (Table 1, col. 2); 2) there is a deterrent effect associated with increases in the audit rate, but in equilibrium it is dominated by the IRS's incentive to audit according to expected yield; and 3) compliance increases with per capita income, but at a decreasing rate, peaking below the maximum per capita income. In addition, there is a significant negative time trend in the audit rate and in compliance (see Table 1, cols. 2 and 4).

This last result appears quite significant; after allowing for a variety of economic and demographic factors and changes over time in the state-level IRS budget per return filed, we still find a significant negative time trend both in the audit rate and in compliance. In an effort to assess the impact of these negative trends on the overall performance of the tax system, we also analyzed the time structure of individual collections per return filed.

Since one finds a deterrent effect of audits in both the 1969 cross-section data set and in the 1977–85 time-series, cross-section data set, the audit rate should be positively related to collections. To test this hypothesis we use individual collections per return (*ICR*) as the dependent variable in a model which is otherwise exactly the same as the one described above. The audit rate again turns out to be endogenous (based on the coefficient of *PIAR* in col. 3 of Table 1), and is, as predicted, positively related to individual collections per return. Surprisingly, there is no significant time trend (see Table 1, col. 3).

The lack of a residual time trend in individual collections per return is surprising and provocative given the negative time trend in audits and noncompliance. A variety of explanations are possible: for example, 1) audits may have become more "efficient" over time and thus have had an increasing deterrent effect, offsetting the decrease in compliance; 2) penalty revisions since 1981 may be improving compliance; 3) shifts in real tax rates over time may have increased collections per return even in the face of declining compliance; or 4) increased use of third-party reports and the "information matching" program may have increased collections per return independent of actual audits.

Further investigation of the time-series data set should help sort out these issues. But already we have learned a great deal. For example, the recent IRS estimates now show the real compliance gap for individuals to have increased from $22 billion in 1978 to $36.8 billion in 1985, a difference of $14.85 billion. If, however, the audit rate had not fallen during this period, our individual-collections-per-return equation indicates that real individual collections would have risen by $15.17 billion in 1985, actually lowering the estimated tax gap in comparison to 1978.

REFERENCES

Allingham, Michael G., and Sandmo, Agnar, "Income Tax Evasion: A Theoretical Analysis," *Journal of Public Economics*, November 1972, *1*, 323–38.

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, March/April 1968, *76*, 169–217.

Border, Kim, and Sobel, Joel, "A Theory of Auditing and Plunder," *Review of Economic Studies*, forthcoming 1987.

Clotfelter, Charles, "Tax Evasion and Tax Rates: An Analysis of Individual Returns," *Review of Economics and Statistics*, August 1983, *65*, 363–73.

Dubin, Jeffrey A. and Wilde, Louis L., "An Empirical Analysis of Federal Income Tax Auditing and Compliance," SSWP No. 615, Caltech, October 1986.

Graetz, Michael J., Reinganum, Jennifer R. and Wilde, Louis L., "The Tax Compliance Game: Toward an Interactive Theory of Law Enforcement," *Journal of Law, Economics, and Organization*, Spring 1986, *2*, 1–32.

Graetz, Michael J. and Wilde, Louis L., "The Economics of Tax Compliance: Fact and Fantasy," *National Tax Journal*, September 1985, *38*, 355–63.

Reinganum, Jennifer F. and Wilde, Louis L., "Income Tax Compliance in a Principal-Agent Framework," *Journal of Public Economics*, January 1985, *26*, 1–18.

Slemrod, Joel, "An Empirical Test for Tax

Evasion," *Review of Economics and Statistics*, May 1985, *67*, 232–38.

Witte, Ann D., and Woodbury, Diane F., "A Test of an Economic Model of Tax Compliance," working paper, Wellesley College, September 1984.

______, **and** ______, "The Effect of Tax Laws and Tax Administration on Tax Compliance: The Case of the U.S. Individual Income Tax," *National Tax Journal*, March 1985, *38*, 1–14.

U.S. Congress, *Tax Compliance Act of 1982 and Related Legislation*, Hearings before the Ways and Means Committee, 97th Cong., 2d Sess., Washington: USGPO, 1982.

GENDER DIFFERENCES IN BEHAVIOR AT HOME AND AT WORK[†]

Gender Differences in the Cost of Displacement: An Empirical Test of Discrimination in the Labor Market

By JANICE FANNING MADDEN*

There are two competing explanations of why women workers earn less than men with equivalent education, work experience, and job tenure: the human capital explanation and the discrimination explanation. The human capital explanation argues that sex differences in human capital investment which arise from sex differences in expectations surrounding labor force participation account for the wage differential. Women workers are expected to invest less in job-specific human capital than otherwise comparable men workers because women expect to spend less time on the job. Furthermore, even for men and women workers with equal *ex post* levels of job tenure and/or work experience, women have invested less in on-the-job training because their a priori expectations of job tenure and/or work experience were less than those of men who now have the same tenure and/or experience. Therefore, in this view, women workers earn less than comparable men because they have invested less in specific human capital. Women earn less because they are less productive; the sex-wage differential is economically efficient.

The discrimination explanation argues that sex differences in labor market opportunities, that is, sex discrimination in the labor market, account for the sex-wage differential. In this view, women workers earn less than comparable men because they are the victims of sex discrimination in the labor market. Women do not earn less because they are less productive; the sex-wage differential is economically inefficient.

While the economic implications of these two explanations of the sex-wage differential are enormously different, both explanations are consistent with empirical studies simply because both resort to "nonmeasurables" to explain the sex-wage differential: empirical studies cannot measure directly either discrimination or job-specific human capital. Therefore, the problem with these two competing explanations of the sex-wage differential is that neither has been empirically sorted from the other. Both explanations are consistent with data which show a wage differential by sex after controlling for education, work experience, and job tenure.

Newly available data on displaced workers provide an opportunity to empirically disentangle these two competing explanations of the sex-wage differential. Displaced workers are workers who have lost their jobs either because their workplaces have closed or because they were permanently laid off due to slack demand for the outputs of their firms. Displaced workers represent a special case of worker mobility. Unlike voluntary job movers, that is, workers who have voluntarily quit their prior jobs, the job mobility of displaced workers is not the result of their own expectations that better jobs are available. A worker who voluntarily changes jobs does so because there is another job which offers higher wages (or other improvements in the conditions of employment). The worker moves precisely because his or her productivity (and wages) is higher on the subsequent job. The wage change is endogenous. Unlike workers who are fired or involuntarily laid off because their personal productivity is lower than that of other

[†] *Discussants*: Rebecca Blank, Princeton University; Elyce Rotella, Indiana University.

*Department of Regional Science, University of Pennsylvania, Philadelphia, PA 19103.

available workers, the lay off of a displaced worker is exogenous to the worker, that is, not the result of his or her individual job performance. The displaced worker, then, offers a unique opportunity for the economist to study the determination of wages and the efficiency of the labor market. For all persons who change jobs, we observe two separate evaluations of the productivity of the same worker. Any comparison of the wages before and after job change of voluntary quitters or of persons fired for cause is complicated by the endogeneity of the wage change. The selection process by which these workers quit or were fired must be explicitly considered. However, for displaced workers who do not choose their status either directly by quitting or indirectly by poor individual job performance, comparisons of these market evaluations are not subject to the endogeneity issues which arise in the study of other job movers.

Changes in the wages of men and women displaced workers before and after their displacement provide an opportunity to empirically distinguish between the two explanations of the sex-wage differential. Workers who have invested more in specific human capital suffer greater wage losses from displacement than workers who have invested less in specific human capital, when wage losses are defined as the difference in the ratio of current-period wages to wages for the period prior to displacement. Displaced workers who have invested in specific human capital suffer greater wage losses because they forfeit returns to any job-specific training which were received on the prior job. For the worker with less job-specific training, the wage on the prior job reflects more general skills. Because general skills have the same influence on productivity and wages on the current job and on the prior job, workers with general human capital suffer less from displacement. If, after controlling for education, experience, and tenure, women workers invest less in specific human capital than men, then women workers are expected to incur less wage losses than men from displacement.

On the other hand, if displaced women workers experience discrimination when they seek subsequent jobs, they choose their next job from a smaller opportunity set than equivalently qualified men who left equivalent jobs. Therefore, women workers who are displaced from jobs which offered the same pay, who engage in the same amount of search, and who are equivalently qualified to their male counterparts are expected to earn lower wages on their subsequent jobs if they are the victims of sex discrimination in the labor market.

The explanation of the sex-wage differential which relies on sex differences in unmeasured investments in job-specific human capital is consistent with women displaced workers experiencing *lesser* wage losses from displacement than their male counterparts. The explanation of the sex-wage differential which relies on sex discrimination in the labor market is consistent with women displaced workers experiencing *greater* wage losses from displacement than their male counterparts. Therefore, an empirical analysis of the effect of gender on the wage losses of displaced workers can provide an empirical test of whether human capital or discrimination accounts for the sex-wage differential.

I. The Data and the General Approach

This study uses data on workers displaced between January 1983 and January 1984 which were collected in the *Displaced Worker Survey* (*DWS*), a special survey of displaced workers who were represented in the January 1984 *Current Population Survey* (*CPS*). The major problem with the *DWS* data is the lack of a control group of nondisplaced workers. However, a control group of "nondisplaced" workers can be constructed by matching the persons surveyed in both the January 1983 *CPS* and the January 1984 *CPS*. All workers over age 19 in 1983 who were living at the same address in 1983 and 1984, who were employed full time in January 1983, who reported their 1983 wages, and who, if employed in January 1984, reported their 1984 wages are included in the data set constructed for this analysis.

The salary losses of displaced workers can be measured by the difference between their wage growth (the ratio of subsequent salary, January 1984, to prior salary, January 1983, *WGLOS*) and the wage growth of nondisplaced workers. *WGLOS* is observed only if the worker is employed in the subsequent period. Therefore, the standard two-stage correction for selection bias as developed by James Heckman (1974) is used here. In the first stage, the probability that an individual employed in the prior period is employed in the subsequent period is estimated to obtain maximum likelihood estimates of the Mill's ratio. Consistent estimates of the determinants of *WGLOS* are obtained in the second-stage *OLS* estimation using the estimates of the Mill's ratio from the first stage as an independent variable. The second-stage equation, estimated including both displaced and nondisplaced workers, is

$$(1) \quad InWGLOS = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i D + \sum_{i=n+1}^{2n} \alpha_i X_i + \alpha_{2n+1} D + \alpha_{2n+2} \lambda + \alpha_{2n+3} \lambda D + \mu$$

where X_i are the i independent variables, discussed in more detail below, which represent worker or job characteristics used in the analysis, D is a dummy variable equal to 1 if the worker was displaced between 1983 and 1984, and λ is the Mill's ratio estimated from the first stage.

Specific human capital is indexed by tenure (*TEN*) and by female representation (*SEXCOMP*) in the prior job. Workers with more tenure are expected to have accumulated more job-specific training. Occupations which employ more women are occupations which have flatter age-earnings profiles. If this occurs because occupations which employ more women require lower rates of investment in job-specific training, then the sex composition of an occupation also indexes the specific human capital investment typical of workers in a given occupation. Workers with more job-specific human capital suffer greater losses from displacement. Therefore, *WGLOS* is lower for displaced workers with more *TEN* and for displaced workers whose prior occupations employed fewer women (i.e., had lower values of *SEXCOMP*). Also, to the extent that specific training is a part of any job, displaced workers who change occupations (*OCCCHG*) or industries (*INDCHG*) suffer greater wage losses than those who are reemployed within the same industry or occupation.

General human capital is indexed by educational attainment (*ED*) and by the total work experience. In the absence of direct work experience measures, age (*AGE*) is an indirect index of work experience. To the extent that general human capital adds to worker productivity on both prior and subsequent jobs, it has no effect on wage losses of displaced workers. However, to the extent that general human capital improves the efficiency with which a worker searches for a job, it increases the wage on the subsequent job, reducing the wage losses of displaced workers.

Discrimination is measured in the customary way. A dummy variable (*FEMALE*) is included to measure the unexplained *WGLOS* associated with the worker being female. (Also, a dummy variable, *BLACK*, is included to measure the unexplained *WGLOS* associated with the worker being black.) If women earn less than men because they have invested less in human capital, then displaced women suffer less of a loss in wage growth than men. If women earn less than men because they face discrimination, then displaced women suffer greater losses in wage growth. However, before this test is performed, it is necessary to consider whether there are other factors which influence the ratio of the subsequent wages to prior wages for displaced workers and, if so, to allow the empirical analysis to control for any gender differences in these factors.

The extent of investment in job search also influences the ratio of subsequent to prior wages for displaced workers. Women are expected to invest less than comparable men in specific human capital because they expect to work for less time. However, if displaced women workers expect shorter work lives than men of similar measurable characteristics, then the gains from job search after displacement are less for women because the present value of future wages on the same next job are less for women. Displaced women of equal prior wages and personal characteristics to displaced men are not expected to have lower search costs than displaced men. Consequently, displaced women are expected to engage in less job search than men resulting in a lower current wage, *ceteris paribus*.

If displaced women have shorter work life expectancies than men, the effects on *WGLOS* of smaller losses in specific human capital could be offset by less investment in job search. Therefore, sex differences in job search must be considered explicitly in the estimation. For displaced workers, *DWS* does report the number of weeks which the displaced worker went without work (*WKSNOJOB*). (These numbers are virtually equivalent for displaced women and men. Men displaced workers went an average of 10.4 weeks without work and women displaced workers went an average of 10.6 weeks without work.) If we include this variable in the regression analysis (which is already controlling for the selection of displaced workers into reemployment with the Mill's ratio), we can control for any sex differences in the intensity of job search which might arise from sex differences in work life expectancies. Then, the coefficient of *FEMALE* can be used to interpret the role of discrimination versus lesser investment in human capital of women workers in determining the gender-wage differential.

If displaced workers were receiving economic rents on their prior jobs, then they suffer a loss in earnings on their subsequent job. Economic rents arise when some factor, such as a union or inefficient management, protects the worker from competition from other workers. If new jobs are not similarly protected, wages are lowered. Because women workers are less likely to be unionized than men and because they are in less powerful unions when they are unionized, the existence of any economic rents for displaced workers on their original jobs results in relatively greater displacement losses for men. Differences in the opportunity to earn economic rents are considered in two ways. First, I control for the location, occupation, and industry of prior job. I control for whether the original job is located in one of the 57 largest SMSAs (*SMSA*) or in the Pacific region (*PAC*), whether the original job is blue collar (*BLUE*) and/or in manufacturing (*MANUF*). Finally, workers who are earning economic rents on their original jobs would be earning higher wages than workers of comparable characteristics. Therefore, if economic rents on prior jobs account for the losses of displaced workers, then workers earning higher *ceteris paribus* wages on their prior jobs (*LNEARNP*) would experience the greatest wage losses.

If workers are compensated for being in jobs with greater risk of displacement, then displaced workers receive lower wages in their next jobs as long as those jobs have lower risks of displacement. In this case, the nominal wages of displaced workers drop, but there is no real wage loss from displacement per se. There is no reason to believe that women or men workers are more likely to be compensated for displacement, although men require greater compensation if they are more likely to continue working beyond the anticipated date of displacement. However, Daniel Hamermesh (1985) cites the steepening of the wage-tenure profiles of displaced workers as the date of displacement approaches as evidence that displacement is a surprise to workers and possibly, to firms. If displacement is not foreseen, then markets cannot compensate workers with higher *ceteris paribus* wages in predisplacement jobs. Therefore, the effect on *WGLOS* of compensation for risk of displacement is ignored in the empirical analysis.

II. Results

When the natural logarithm of *WGLOS*, the ratio of weekly wage in January 1984 to weekly wage in January 1983, is regressed on gender and all the other variables described above except for *WKSNOJOB* and *SEXCOMP* as indicated by equation (1), the coefficient for *FEMALE* is $-.106$ with a standard error of .066, indicating that women displaced workers of equivalent age, education, industry, occupation, location and wage in 1983 to displaced men workers experience a 10.6 percent greater loss in wage growth between 1983 and 1984. When a similar specification is estimated separately by sex but the losses from displacement are indicated by a dummy variable taking the value of 1 when the worker is displaced between 1983 and 1984, the coefficient for the displacement dummy variable is −.158 with a standard error of .032 for men and −.271 with a standard error of .049 for women. Therefore, under both approaches, displaced women lose about 11 percent more salary growth than men.

While the lower wage growth of women displaced workers is consistent with the sex discrimination in the labor market explanation of the gender wage differential, this equation does not take account of potential gender differences in investment in job search which could arise from gender differences in work life expectancies. When the variable *WKSNOJOB* is added to the estimated equation, the coefficient on *FEMALE* increases, as expected, but remains negative at −.095 with a standard error of .062.

When *SEXCOMP* is added to the equation, the coefficient on *FEMALE* decreases to −.133 with a standard error of .066. The coefficient on *SEXCOMP* is positive (but insignificant), indicating that workers displaced from occupations which employ relatively more females suffer less of a decrease in wage growth than those displaced from jobs in occupations which employ more men. The sign of the coefficient of *SEXCOMP* is expected if female occupations require smaller investments in job-specific human capital. However, the finding that controlling for the sex composition of prior occupation increases the estimated loss from displacement for women relative to men merits some further discussion.

Both the human capital explanation of the gender-wage differential and the discrimination explanation based on comparable worth arguments are consistent with workers displaced from female jobs experiencing less loss of wage growth than workers displaced from male jobs. The comparable worth advocate argues that jobs which are disproportionately female underpay their occupants relative to jobs that are disproportionately male, *ceteris paribus*. Workers displaced from female jobs do not suffer as much wage loss because they were underpaid in those jobs and are therefore more likely to find a wage improvement on their next job than are workers displaced from male jobs. In order to empirically sort these two competing explanations of the effect of sex composition of job on wage losses from displacement, it is necessary to consider the effect of sex composition of *subsequent* job on wage losses of displaced workers. Under the human capital explanation, if displaced workers move on to jobs which employ more females, they are entering jobs which require less of an investment in specific human capital. *Ceteris paribus*, jobs requiring less investment in job-specific human capital start at higher wage rates. Therefore, displaced workers who move into more-female jobs following their displacement are expected to suffer less wage loss (when loss is measured as the difference between salary at time of displacement and the salary at start of next job) than those moving into more-male jobs. Under the comparable worth explanation of wage differentials between male jobs and female jobs, jobs which include relatively more female workers pay lower wages, *ceteris paribus*. Therefore, displaced workers who move into more-female jobs following their displacement are expected to suffer greater wage losses than those moving into more-male jobs. Consequently, we have a clear test of comparable worth vs. specific human capital investment explanations of wage differentials between male- and female-dominated jobs.

The results of estimating the version of equation (1) which includes the most comprehensive list of independent variables are listed in Table 1. Specifically, this version adds a variable reflecting the sex composition of the 1984 job as well as that of the 1983 job. The coefficient on *SEXCOMP* for the 1984 job is a statistically significant −.004. Separate estimation of this equation for men and for women yields equivalent coefficients for *SECOMP84*. The negative sign is consistent with the comparable worth

TABLE 1—DETERMINANTS OF ln*WGLOS*: Coefficients for Displaced Workers Only

Independent Variable	Coefficient	Standard Error
FEMALE	−.040	.069
BLACK	−.054	.089
TEN	−.012	.004
AGE	.073	.029
AGE^2	−.0009	.0004
ED	.017	.020
SMSA	−.084	.059
PAC	.028	.071
BLUE	−.104	.071
INDCHG	−.155	.053
OCCCHG	.033	.057
LNEARN83	.034	.055
WKSNOJOB	−.007	.002
SEXCOMP83	.003	.001
SEXCOMP84	−.004	.001
Mill's Ratio	−2.016	1.487
N	3217	
R^2 (adj.)	.19	
MSE	.11	

Note: The data set used in estimating this equation includes both displaced and nondisplaced workers. The equation also included estimates of these same coefficients for all workers, regardless of displacement status. Therefore, the reported coefficients indicate the net effect of these characteristics on ln*WGLOS* for displaced workers.

advocate's expectation and inconsistent with the expectation of human capital theory. (Of course, these results are also consistent with other, "nonhuman capital," theories of occupational wage differentials.) Furthermore, the coefficient on *FEMALE* has increased to −.040. The inclusion of the sex composition of the subsequent job substantially reduces the estimate of the wage growth loss of displaced women relative to displaced men. This result is consistent with sex discrimination in the labor market limiting the job alternatives for women, particularly in those jobs which employ relatively fewer women.

The estimated coefficients for the other independent variables in equation (1), as listed in Table 1, are consistent with expectations. As expected by human capital theory, *TEN* and *INDCHG* are significantly negative, indicating that displaced workers who change industry or who have more tenure on their original jobs suffer greater losses in wage growth. Also, workers with general human capital as indicated by *ED* and *AGE* suffer significantly less loss in wage growth. Consistent with the premise that displaced workers who suffer from labor market discrimination would experience less wage growth, the coefficient for *BLACK* is also negative.

III. Conclusions

The wage effects of the job mobility of displaced workers provide a unique opportunity to test whether the sex differential in wages can be attributed to unobserved differences in human capital investment by sex which arise from different expectations of lifetime labor force participation or to sex discrimination in the labor market. Because the wage change induced by displacement is exogenous in that the observed mobility is not itself the result of the expected wage change, differences in the observed wage change for male and female displaced workers reflect the characteristics of the labor market and of the jobs involved, rather than the personal characteristics of the workers. If women have invested less than comparable men in job-specific human capital on the jobs from which they were displaced, then women are expected to suffer lesser wage losses from displacement. If women face a labor market which offers them less opportunity, then women are expected to suffer greater wage losses from displacement when they are displaced from similar jobs. For workers displaced between January 1983 and January 1984, women experienced a greater wage loss than men. Therefore, the evidence from displaced workers suggests that the sex-wage differential arises from discrimination in the labor market and not from unobserved lower rates of investment by women in on the job training.

REFERENCES

Hamermesh, Daniel, "The Costs of Worker Displacement," NBER Working Paper, December 1985.

Heckman, James, "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

Gender, Unions, and Internal Labor Markets: Evidence from the Public Sector in Two States

By DEBORAH M. FIGART*

Economists have been measuring, developing theories for, and explaining the sex-based wage differential for several decades. Some theories, such as human capital theory, have focused exclusively on the characteristics and progress of individuals. Institutional approaches, such as dual labor market and occupational sex segregation theories, look solely at the structure of the labor market. However, examining average earnings by sex, occupational titles, or sectors masks another fundamental cause of the wage gap: the lack of career advancement by women in both female- and male-dominated occupations. The reason is that women are on a lower or different occupational ladder, what can be called an internal labor market, than men. This paper combines the insights of earlier approaches by highlighting the effect that internal job structures and career ladders have in shaping individuals' opportunities. The hypothesis is that unions play a major role in altering the structure of opportunities, creating ladders, and allowing women access to higher-level jobs, as they have done for men.

I. Developing a Theoretical Framework for Examining Equity

It is argued that inequitable outcomes result from gender-based stereotypes and expectations, which produce differences in institutional structures faced by women in the labor market. The theoretical foundation that best explains this picture is internal labor market theory, redefined for gender. The internal labor market distributes positions within a particular workplace (see Peter Doeringer and Michael Piore, 1971). Many of the rules and procedures of the internal labor market were created and practiced in large, unionized plants. Although unionized workplaces are not the only ones with hierarchical ladders, unions have created many clearly defined careers in an internal labor market for men. Firms and employers have restructured work and/or introduced technological change to raise productivity in conjunction with union-negotiated pay increases. Unions have also negotiated on-the-job training and well-defined systems of progression (Richard Freeman and James Medoff, 1984).

Unlike in many male-dominated occupations, in female-dominated occupations there are not adequate ladders, and the few administrative positions are held disproportionately by men. If it is true that women are clumped at the bottom of the earnings distribution within an internal labor market, then the way to achieve pay equity is to move them up the occupational ladder, or to develop ladders where they do not exist. A comparable worth strategy based on raising entry-level or average wages will not by itself lead to pay equity. The second step necessary is job restructuring and building adequate career ladders for women.

Thus far in the 1980's, union membership has given women a wage boost, often more than what men have gained. BLS data demonstrate that the sex-based wage gap throughout the life cycle (until retirement) is less for union members. Similarly, the ratio of union to nonunion wage rates for both sexes is greater than one, with women's ratios greater than men's ratios. The largest wage boosts for union women relative to nonunion women are in the earnings years 35 to 44 and 45 to 54, precisely when women lose in terms of fewer mobility opportunities vis-à-vis men.

*Assistant Professor of Economics, Washington Semester Program, American University, Washington, D.C. 20016.

II. Estimating Career Ladder Differences by Sex and Union Status

This paper goes one step further than the empirical work on how sex and unionization affect wages and analyzes the differences in mobility opportunities in the public sector in two states. New York State, with a comprehensive collective bargaining law, is compared with Maryland, a state with no statewide collective bargaining. The unions considered are the Civil Service Employees Association (CSEA), AFSCME (representing paraprofessional and office/clerical occupations), and the Public Employees Federation, SEIU (representing technical and professional occupations). The craft and service occupations in New York are also covered by collective bargaining, but most officials/administrators are not. The data (from Equal Employment Opportunity Commission Form 164 or EEO-4 files in 1975, 1980, and 1984) are grouped, comparing male- and female-dominated occupations in both states. Rather than surveying individuals, a "snapshot" is presented of 1) the structure of job opportunities by sex within states, to view sex discrimination in ladders; and 2) the structure of job opportunities between states, to view the impact of unions on career ladders.

First, a proxy for the length and depth of career ladders was developed. Statistical tests assessing the different mobility opportunities for men and women in the two states were performed utilizing the proxy. Second, structural analysis compared sample ladders by state and sex.

To create the proxy, the annual salary of full-time men vs. women or of New York vs. Maryland employees was compared. Specifically, a measure was used denoting the accumulated probability of being at a higher annual salary for each of eight salary ranges: $100-(\text{Prob}\,\Sigma_{i=1}^{8} i)$ where i is the accumulated percentage of the population (sex or state group in the seven occupations) at each salary range. The "mobility probability" values at the first salary range are close to 100 percent and the values decline as salary rises, reaching zero at the eighth or top range. By connecting the mobility probability values, a Mobility Probability Line (MPL) is constructed which has a negative slope. Since all the values represent 100 percent minus the accumulated percent of individuals at each salary level, we can think of them conceptually as a proxy for the probability of "moving ahead to" or "being at" a higher salary level.

First, it was expected that within the states, men's MPLs are significantly greater than women's MPLs. That is, because women are on low opportunity ladders, the probability of advancement for women falls more quickly as salary increases. Second, the difference in mobility values between men and women, which is called the "sex gap," was expected to be greater in the nonunion state (Maryland) than the union state (New York). Lastly, it was hypothesized that unions in New York have increased the likelihood of women's and all employees' career advancement relative to their Maryland counterparts for all occupations and years. The following groups were tested corresponding to the hypotheses: 1) men vs. women within each state (sex differences); 2) unionized women and men vs. nonunionized women and men between each state (state differences by sex); and 3) all union employees versus their nonunion counterparts (state differences).

Two sets of regression analysis are performed using the OLS method. The first regression set analyzed the sex gap and the second set analyzed the height of the MPL and the state gap. Overall, the sex gap measures the degree to which equity exists in each state or occupation. The height of the MPL is a general measure of employment/economic status. The regression equations are

$$(1)\quad SEXGAP_k = a_0 + a_i OCC_j + a_7 Y1975 + a_8 Y1984 + a_9 STATE + a_{10} SALARY + u$$

$$(2)\quad MPV_k = B_0 + B_i OCC_j + B_7 Y1980 + B_8 Y1984 + B_9 SEX + B_{10} STATE + B_{11} SALARY + u$$

where $SEXGAP_k$ is the absolute difference in men's minus women's mobility probability values at each of k or eight salary levels; *STATE* is 1 if New York; *SEX* is 1 if male; *Y1975*, *Y1980*, and *Y1984* are 1 in those years, respectively; *SALARY* is midpoint values for the ranges in the EEO–4 data; and the OCC_j's are 1 in official/administrator, professional, paraprofessional, craft, clerical, or service occupations respectively.

Additionally, pairwise Chow tests were performed on the models using all eight, plus the middle six and four salary levels (see Maryellen Kelley, 1982, for an example of this procedure applied to seniority differentials). The reason for this breakdown is that very few employees are distributed at the upper and lower "tails" of the salary ranges. One can also hypothesize that upward mobility possibilities are created or destroyed at middle salary ranges. Five pairwise Chow tests were performed to see if mobility probability values across salary levels, or the MPL, for the following groups are statistically different: 1) New York men and Maryland men; 2) New York women and Maryland women; 3) New York employees and Maryland employees; 4) women in New York and Maryland; and 5) men in New York and Maryland.

To create sample ladders in each state for typical female- and male-dominated occupations, a series of jobs within several occupational groups was selected which listed the job titles from lowest to highest grade and their respective salaries. The job titles were chosen by looking at the New York and Maryland occupational codes (principally the first four to five digits of a seven-digit number).

III. Findings

The importance of occupational sex segregation on women's mobility probability is borne out by the regression results of equation (1). In both states the sex gap is greater in male-dominated than in female-dominated occupations. The occupation which ranks as having the greatest sex gap is the skilled crafts. The occupation which has the least is paraprofessionals. In between these, the male-dominated occupations generally have the greatest mobility probability gaps by sex. Substituting percent female of the occupation in each year by state for the occupational dummies also demonstrates that as the percent female in the occupation rises, the sex gap declines significantly. In fact, the only variable significant in the additional regression is percent female.

These results are consistent with earlier feminist studies demonstrating that as percent female in the occupation rises, median wage levels fall. However, occupation is not the only determinant of mobility probability. It is equally important to analyze the gap within occupations. The result show that in each of the three years, in every occupation, men's MPLs are greater than women's MPLs.

Absolute changes in the percentage differences in mobility probability values depict the dynamics of the sex gap over the decade from 1975 to 1984. The sex gap declined, more in New York than Maryland. Hence, one might think that over time the wage gap and mobility probability differences between women and men have diminished and will disappear in time. This is certainly argued by some opponents of pay equity legislation. However, at the highest salary levels, the sex gap has worsened over the decade. Thus, women are moving up the middle ranks, but not the top ranks.

The results of the first regression also show that the sex gap by state is significant, but in the opposite direction than expected. The coefficient for the state (union) dummy has a highly significant positive value, 4.38. This means that the sex gap in New York is significantly greater than the sex gap in Maryland, necessitating a rejection of the second hypothesis.

But we cannot conclude that unionization does not positively affect women's status in internal labor markets in absolute terms because the height of the MPLs by state are different. The results of this second regression are presented in Table 1. As shown earlier, "being male" clearly exhibits a significantly positive effect on mobility probability values. This is especially true in New York, substantiating the fact that the sex gap is significant. Though not significant, working in New York has a positive effect on

TABLE 1—MOBILITY PROBABILITY EQUATIONS[a]

Variable	Maryland	New York	Both States
Y1980	15.52[e]	8.55[d]	12.04[e]
	(6.00)	(3.30)	(6.53)
Y1984	30.97[e]	35.55[e]	33.26[e]
	(11.53)	(13.26)	(17.41)
Sex	5.01[c]	9.39[e]	7.17[e]
	(2.37)	(4.43)	(4.76)
State	–[b]	–[b]	1.56
			(1.04)
Off/Adm	21.34[e]	16.84[e]	19.09[e]
	(5.40)	(4.25)	(6.78)
Prof	12.44[d]	12.61[e]	12.53[e]
	(3.15)	(3.19)	(4.45)
Para	−12.18[c]	−7.71[c]	−9.94[e]
	(−3.08)	(−1.95)	(−3.53)
Craft	−3.49	−7.28	−5.40
	(−.88)	(−1.84)	(−1.92)
Clerical	−10.92	−11.53[c]	−11.20[e]
	(−2.77)	(−2.91)	(−3.98)
Service	−16.76	−10.87	−13.81[e]
	(−4.24)	(−2.74)	(−4.91)
Salary	−.004[e]	−.004[e]	−.004[e]
	(−34.10)	(−33.23)	(−47.31)
Intercept	96.23[e]	94.75[e]	94.75[e]
	(26.26)	(25.71)	(34.78)
R^2	.81	.79	.79
Number in Sample	336	336	672

[a] The t-statistics are shown in parentheses.

[b] Not a variable in the regression.

[c], [d], and [e] mean significant at the .05, .01, and .001 levels, respectively.

TABLE 2—CHOW TEST USING FOUR SALARY LEVELS[a]

Pool	Subgroups	F Ratio	df_2
States	NY, MD	6.15[c]	314
New York	Men, Women	10.73[c]	148
Maryland	Men, Women	2.02[b]	148
Men	NY, MD	4.79[c]	148
Women	NY, MD	3.57[c]	148

[a] df_2 is the degrees of freedom in the denominator of the F-test;

[b],[c] means reject that the subgroups are identical at the .05 or .01 levels, respectively.

mobility probability for both sexes, especially men.

The state dummy does become significant when the analysis focuses on the middle salary levels. The importance of adding state differences to the analysis is that overall, both New York men and women do better in terms of mobility probability than Maryland men and women. But the gain is greatest for men. Thus it appears that unions do increase opportunities for career advancement for all workers; however, they increase men's opportunities more than women's. This is perhaps why I obtained the unexpected result that the sex gap was greater in the union state.

The state gap in mobility probability also increased over the decade, most remarkably in the female-dominated paraprofessions, possibly indicating that unionization benefits women in a two-stage process. In the first stage, immediate, sizeable gains in career opportunities (mobility probability) accentuate the preexisting sex gap. In the second stage, when overall wage gains are less sizeable and specific policies target women's mobility, women may be closing the gap between themselves and their union brothers.

The results of the Chow tests corroborate the preceding findings—especially in the models focusing on the middle salary levels. Significant differences are evident in the treatment of each of the five subgroups tested. Table 2 presents the results for the Chow test on the four middle salary levels.

The second part of the analysis demonstrates the institutional factors underlying the quantitative results. Mobility probability is different between the sexes and states because individual workers face different opportunity structures. The sample career ladders developed indicate that it is these structural reasons that determine career mobility and ultimately wages over the life cycle.

For example, men's ladders are longer in selected occupations in New York and Maryland. Male technicians (in electronics and drafting, for instance) have ladders with more rungs, and higher entry and peak grades/salaries then female account clerks or library assistants. Similarly, in New York, payroll audit clerks have ladders with five rungs, from grades 5 through 22, with crossovers to audit clerks at various grades. Maryland's central payroll clerks have ladders consisting of only four rungs, beginning at grade 6 and ending at grade 11.

To demonstrate that unionization played a significant role in accounting for the state differences, the extent of unionization in each state and the policies and programs gained by New York employees via collective bargaining were analyzed. A comparison of the two civil service systems indicates that they share similar principles and procedures for hiring and promotion, and the occupational profiles of women in the states are similar. The main difference is the existence of state-wide bargaining in New York.

Collective bargaining agreements between the state of New York and CSEA contain provisions for developing joint labor-management committees to study, develop, implement, and monitor career advancement through career development programs. One program, the Clerical and Secretarial Employee Advancement Program (CSEAP) was negotiated in the 1979 contract between the three large CSEA units and New York State. CSEAP provided $1.9 million in funds for increased education and training for members and created a mechanism for building bridges into career ladders in many female-dominated occupations through job restructuring, transitional exams, and upgrades with technological change. Many proposals target "dead-end" jobs to help eliminate the absence of a clear line of progression from one job to another. In contrast, there are no such career development programs in Maryland. Therefore, while it is likely that unions are not the only ones responsible for the differences between the states or the advances in New York, it is clear that they have some role.

IV. Conclusion

The empirical results show a significant mobility probability gap by sex within occupations and differences in mobility probability between occupations. It appears that once on the job, it is the structure of the internal labor market, not solely human capital characteristics or labor force commitment, that determines position and status. The significant difference in mobility probability values by state suggests that unions have an impact on earnings and the structure of the internal labor market. Unfortunately, the benefits of unionization accrue, at least in the short run, mostly to men. Hopefully, in a second stage, women can press for strategies such as job restructuring or other career development programs that can narrow the sex gap in mobility probability. As a whole, the results suggest that comparable worth and improved career ladders work hand in hand. Thus, policy analysts searching for sources of inequity in the labor market need to address the "ladder problem" for women, as well as wage discrimination and classification bias. Pay equity, not just at the starting gate but at the finish line, could be the result.

REFERENCES

Doeringer, Peter B. and Piore, Michael J., *Internal Labor Markets and Manpower Analysis*, Lexington: D.C. Heath, 1971.

Freeman, Richard B. and Medoff, James L., *What Do Unions Do*?, New York: Basic Books, 1984.

Kelley, Maryellen R., "Discrimination in Seniority Systems: A Case Study," *Industrial and Labor Relations Review*, October 1982, *36*, 40–54.

U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings*, Washington: USGPO, January 1986.

Nonprofit Firms in Medical Markets

By MARK V. PAULY*

The medical care industry is characterized by a large market share of output produced in firms which are organized on a not-for-profit basis. The market share of not-for-profit acute care hospitals is especially high, with such firms accounting for 70.3 percent of all beds in 1984, but these firms also exist in nursing homes, in other long-term care, in home health, in out-patient dialysis centers, and in most other parts of the health care industry.

The fact that a firm is not organized with the explicit goal of maximizing profits or stockholders' wealth is, in itself, a reason to be skeptical about the appropriateness of applying the conventional neoclassical models of firms and of markets without qualification. But what sort of qualification is appropriate, and how important is it to make adjustments? Since nonprofit firms selling services directly to consumers characterize other markets as well, the types of answers to this question which have been generated in medical economics provide insights beyond medical care. More importantly, the theoretical and empirical investigation of behavior in such atypical firms itself sheds light on the impact of firm organizational structure and within-firm incentives on firm and market performance in the for-profit sector. There are messages here for the "new industrial organization," as described by Oliver Williamson (1975).

I. Property Rights Differences Between For-Profit and Nonprofit Firms

There are three major differences in the institutional constraints facing a not-for-profit firm, as compared to the neoclassical for-profit firm. First, not-for-profit firms must look to donations for initial equity capital; they do not have the power to obtain capital in return for the promise of a share of the residual income of the firm. Second, not-for-profit firms are not permitted to pay out as cash dividends any revenues in excess of production costs and cost of debt; residual returns are not alienable. Legal rules even inhibit the ability of managers of the firm to add profits to their salaries *ex post*. Third, not-for-profit firms cannot be sold or liquidated for proceeds to be paid to a set of individual owners.

These institutional differences in the right to transfer wealth have potential consequences for the incentives faced by those who direct the firm. The most obvious difference, and the one that has been subject to the most discussion, is that the "decision makers" who are unable to extract residual income in the form of cash (because of a kind of an attenuation of property rights) will choose to take it in other forms. A related notion is that the inability to sell ownership shares in the firm may lead to a difference in the ability of the firm to obtain large amounts of additional equity capital for capital investment. Differences in tax treatment and requirements to furnish charity care are also important, but will not be discussed here.

[†] *Discussants*: Burton Weisbrod, University of Wisconsin-Madison; Charles E. Phelps, University of Rochester; Roger D. Feldman, University of Minnesota.

*Robert D. Eilers Professor of Health Care Management and Economics, Professor of Public Management and Economics, University of Pennsylvania, and Executive Director, Leonard Davis Institute of Health Economics, 3641 Locust Walk, Philadelphia, PA 19104–6218.

II. Models of Nonprofit Hospitals

Most of my discussion will refer to hospital ownership structure. There are two, so far rather distinct, strands of analysis in the discussion of the economics of nonprofit hospitals. One set of models has assumed that equity capital has already been obtained, ignored philanthropic motivation, postulated various sets of objectives for the firm, and carefully modeled the maximization of a utility function in those objectives subject to a break-even constraint. Another approach has modeled the role of voluntary donations in the establishment of nonprofit enterprises, but with only a rudimentary behavioral model of production in the enterprise. There is as yet no model which fully integrates the two approaches.

The "objective function" models all postulate an exogenous break-even (or maximum deficit) constraint, but differ in terms of the objectives the hospital is thought to pursue. One set of models postulates that the hospital seeks to maximize the money incomes of a set of decisive agents, particularly the physicians on the hospital's medical staff. Another set of models simply modifies the form in which a nonprofit hospital can pay out profits. The hospital still prices and/or chooses outputs and inputs as would a profit-maximizing firm, but pays "dividends-in-kind" to decision makers (managers or trustees). As Patricia Danzon has noted, "Although the rights to residual profit in a non-profit hospital are not well-defined, profit maximization is nevertheless an appropriate model provided the various claimants can agree on maximizing their joint gain" (1982, p. 38). The third class of models provides an exception to Danzon's conclusion—if output itself yields utility to decision makers, or if something which affects demand (say, "quality") also yields utility to hospital decision makers, then obviously the profit-maximizing price, quality, or output need not be selected.

There are many behavioral models, but little consensus on the appropriate one. One reason for this failure to achieve consensus has to do with the difficulty of distinguishing empirically among theoretically plausible models. The presence of profits in the constraint means that all of the variables which affect profits appear in the comparative statics of each of these models, as, of course, they appear in those of the profit-maximization model. Since the same variables with the same predicted signs show up in all models, it is obviously impossible to distinguish among them on this basis. The only real difference among models is that some include variables which others do not explicitly include. For example, the physician income-maximization model points to dimensions of physician productivity and pricing which may affect what the hospital does. But since other models do not explicitly rule out such influences, no model can be refuted. The heart of the theoretical problem is that the objective function is usually unobservable. Utility always is unobservable, and direct test of whether profit or physician income is at its maximum is usually impossible.

With such "internal" tests known to be inconclusive, the alternative empirical strategy is to compare not-for-profit hospitals, with unknown objectives, to for-profit hospitals whose objectives presumably are known. Detecting differences in behavior and evaluating these differences should shed light on differences in objectives. But what differences should one expect? The physician income-maximization model of the not-for-profit firm implies that, with a given number of medical staff members, the hospital chooses the same levels of output and quality as it would if it maximized money profit. The total price (hospital price plus physician price) is also the same in such a model. Only the division of that price between payments received by physicians and payments received by the hospital potentially differs from the profit-maximization case. The second class of models mentioned above also implies no difference in price, quality, or output between for profit and not for profit firms. Cost would differ, but only because of inclusion of the utility-generating dividends-in-kind in total cost. Only the third model suggest a difference in important aspects of equilibrium. Hence, comparisons of behavior across firm types will display differences only if the third type of model is appropriate, and a finding of no difference would not be surprising.

Much of the motivation for the debate concerning expected differences has concerned the alleged "inefficiency" of the not-for-profit firm because of the attenuation of property rights. Even at a theoretical level, this basic question has not been fully resolved, largely, I believe, because inefficiency in production, in the sense of deviation of resource costs from their theoretical minimum for a given amount and type of output, does not have a well-defined role in models of not-for-profit firms and plays no role whatever in models of for-profit firms.

The inefficiency associated with not-for-profit ownership could be considered to be analogous to a tax. In the conventional analysis of the impact of a pure profits tax on a for-profit firm, the partial-equilibrium conclusion is that such a tax simply reduces rents; it does not cause the firm to deviate from efficient production. One way to look at the "no-cash-payment" rule under which not-for-profit firms must operate is as a kind of inefficient tax, one requiring profits to be paid out in a form which may yield less utility than if they were paid in cash. Particularly in the type of model Danzon outlined, such dividends show up as costs, not as profits. But there is no implication here that there will be inefficiency in production per se. Output is produced with minimum levels of all inputs, and other "costs" are just the accounting implications of dividends-in-kind. Technically, the attenuation of property rights does not lead to inefficiency in production, only to payment of dividends in an inefficient way.

Where characteristics of output, such as quality or volume, yield utility directly, these dividends-in-kind may not be so easily segregated, even conceptually. Nevertheless, none of the utility-maximization models lead to predictions of technically inefficient production, given quantity or quality, and the excessive quality or quantity from the viewpoint of demanders of outputs is just the mirror image of the form of dividends-in-kind.

The real inefficiency therefore arises to the extent that dividends-in-kind are valued less highly by their recipients than would be the cash their cost represents. In the physician-dominated model, at least in a world of certainty and absent administrative costs, there is no inefficiency of this kind. And even in the other models, there would potentially be offsets in cash wages for managers who like "quality," who would choose higher quality as part of their real income even in the profit maximizing form. There is inefficiency only if such managers are overly compensated at the margin in the form of quality or technology.

A test of differences arising from ownership is also made problematic in part because firms of different ownership structures actually do coexist in the same market. Are there certain types of outputs for which certain types of firms are more suitable? The second strand of analysis relates the existence of such firms to certain characteristics of output. Henry Hansmann (1980) has outlined part of the argument. Suppose the quality of output is a positive argument in a nonprofit firm's utility function. Suppose also that some dimensions of quality cannot be observed well by consumers, and that they are aware of this difficulty. Then consumers may prefer to buy such output from not-for-profit firms. In effect, the consumer "sees through" the economic behavior model, identifies the differences in incentives within each type of firm, and chooses that type of firm whose estimated equilibrium value of the unobserved quality level comes closer to the consumer's most preferred option. Neither type of firm is more efficient in some intrinsic sense; each is preferred for a particular type of output or consumer, with the not-for-profit firm possibly sacrificing some productive efficiency in order to assure a quality level closer to what people want. Even this sacrifice need not occur if some managers have a strong enough taste for quality. The not-for-profit form itself is a signal to consumers about which types of firms have managers with a stronger interest in quality.

Given that hospitals of different ownership types can coexist, what would be the expected impact of entry on long-run equilibrium? Obviously competition will constrain the rents that dominant groups can earn, whether these groups are investors, medical staff, or administration and trustees. Morever, while greater competition or other financial pressures will reduce the extent to which members of such groups can get what they want, one should not confuse such external constraints with a change in objectives. For instance, greater competition will compel individual physicians in a physician dominated hospital to be constrained by the medical staff as a whole—in the collective best interest.

Free entry will, as Joseph Newhouse (1970) noted some time ago, force not-for-profit firms to price in long-run equilibrium at the same level as would for profit firms. Other objectives can still be pursued, but only as far as costs equivalent to normal profit on equity capital will allow. The product differentation discussed above will also persist in long-run equilibrium, except that the not-for-profit firm will be forced to produce efficiently whatever quality it chooses.

III. Donations, Altruism, and Nonprofit Hospital Objectives

Another aspect of the second type of theory is related to the observation that initial equity capital for nonprofit hospitals typically comes from philanthropic donations. There are two issues to be noted here. First, I will review theories of donation, and attempt to integrate them with theories of production. Then I will examine the role of donated equity, and of private not-for-profit hospitals as a whole.

Burton Weisbrod (1977) has provided the most comprehensive theory of nonprofit donations. He argues that donations occur when the government fails to provide collective goods which at least some citizens value more than their cost. Deviations of tax bases from marginal benefit taxes would provide the basis for such donations. The emergence of a not-for-profit *firm* as a recipient of donations can be explained by an argument analogous to Hansmann's explanation of nonprofit output. If donors have difficulty monitoring the quality or quantity of output, or the price or profit levels set by the firm, donations to for-profit firms can be absorbed into profits (see Eugene Fama and Michael Jensen, 1983). In contrast, nonprofit firms may be more likely to use donations for their intended purposes, particularly if the objectives of the firm are consistent with those of the donors.

The quality-monitoring aspects even extend to the physician-dominated form. Surely physicians are in the best position to monitor quality within the hospital, and surely raising hospital capital via donations will, up to a point, be less costly than tapping the conventional capital market. By deviating somewhat from short-run maximization of physician income, and channeling donations to their intended recipients, physicians may be able to attract donations for their hospital. Physicians may also be likely to count hospital quality as part of their real income. Their position in the "trust market" may make them especially suited to be monitors of the use of donations; donors and consumers may then be content to monitor the monitors, and to rely on short-run impediments to profit maximization as a way of constraining the medical staff. It is quite possible that such a strategy may yield larger long-run profits to physicians than would pure profit maximization.

IV. Empirical Evidence: The Time-Series

The externality-philanthropy view of the not-for-profit hospital provides an interesting perspective on the recent history of such an organizational form. Many people in the medical care industry have bemoaned the recent growth in the share of investor-owned for-profit hospitals (see Arnold Relman, 1980). The actual growth has been less than spectacular. The share of beds in investor-owned hospitals grew, not at the expense of the private nonprofit hospital, but almost entirely by a shrinkage in the share of local government-owned hospitals, and has only increased from 8 percent of beds in 1975 to 10 percent in 1984.

This period corresponds to one in which the relative share of federal government-financed care grew rapidly. In effect, financing of care for low-income or uninsured elderly people was transferred from the voluntary sector to the government. In such an environment it should not be surprising that investor-owned enterprises should grow, especially in expanding markets dominated by cost-based payment. In such markets, it was difficult to mobilize support for new philanthropy when, in principle and in plan, all of the indigent and the medically indigent were to become clients of the government. Thus the relative growth of for-profit firms was a direct response to the erosion of the supply of philanthropic capital. Not-for-profit firms found it difficult to expand or to start, and for-profit firms were able to locate where positive profits could be generated. In this sense, the modest growth of for-profit firms was caused, not by an upsurge of venality, but rather by a substitution of tax-financed charity and private production for the previous vertically integrated solicitation-production combination represented by the not-for-profit firm.

V. Empirical Evidence: The Cross Section

As noted above, the test of theories of nonprofit firms has to be empirical. There have been many such comparisons. Although they differ in detail, the overall impression is that there is little ownership-related difference in hospital cost given quality, or in quality given bed size, teaching status, and other proxies for type of output. The evidence for nursing homes is less clear. There is obvious specialization by nonprofit nursing home firms in high-quality output, and some evidence that the for-profit firms disappoint those consumers who are poorly informed, as Weisbrod and Mark Schlesinger (1985) have noted.

The one area in which there may be differences between for-profit and not-for-profit hospitals is in price levels. A recent extensive study (Michael Watt et al., 1986) claims to have found differences in charge levels of approximately 17 percent for matched sets of for-profit and not-for-profit hospitals. These differences are reduced to about 10 percent when differences in tax burdens are netted out.

My own impression is that a difference of this magnitude, especially given the failure of other studies (for instance, Frank Sloan and Robert Vraciu, 1983) to find any differential, is not particularly strong evidence for important consequences arising from ownership structure. Since for-profit firms can probably shift equity capital more easily, the main message may simply be that such firms are more likely to be found in markets in which any firm, for-profit or not-for-profit, would charge high prices. There is an important message here: location and production by for-profit firms is endogenous. But no empirical study has taken account of the endogenous nature of ownership structure. This failure renders suspect all of the existing studies attempting to compare prices, profits, and costs between ownership types. Since endogenous for-profit firms are more likely to locate where market conditions permit high profits and prices for any firm, for-profit or not, there is a bias toward finding higher prices in for-profit firms.

One clear message from recent work is that, despite the anomalous character of the not-for-profit form, theory does not predict wide differences in behavior at the level of the market, nor does empirical evidence suggest that large differences do occur. There is however, the potential for wide variations in market structure as measured by ownership shares, even if there is relatively little difference in efficiency, cost, or welfare associated with different forms. Especially if entry is free and consumers somewhat discriminating, relatively small changes in advantages for one organizational form over another can lead to large changes in the share of output produced by not-for-profit relative to for-profit firms. The growth of investor-owned hospitals may be an example of this phenomenon, as is the growth of multi-institutional affiliation in the health care industry.

VI. The Welfare Economics of Nonprofit Firms

The other major message is that, with free entry and markets that adjust in other ways, there need not be inefficiency created by deviations in ownership forms from the conventional profit-maximizing one. In particular, even ignoring the gains from providing poorly informed consumers with quality closer to their preferred levels, as long as there are sufficiently many managers who place sufficient subjective value on the quality of the output they supervise, there need be no net inefficiency arising from the dividends-in-kind feature of not-for-profit firms. Moreover, gains to poorly informed consumers, who know that they are ignorant and know that the not-for-profit form is better for them, can offset any inefficiency cost even if there are excessive dividends-in-kind. The not-for-profit form may be the second best optimal way of dealing with asymmetric information, exactly as Kenneth Arrow (1963) has suggested. The interesting question is whether this equilibrium with nonprofit firms is (second-best) optimal. Is there an invisible hand which brings such forms into existence when (and only when) they are needed?

There are some potential impediments. The differential tax treatment of for-profit and nonprofit firms may itself cause excessive spread of the not-for-profit form. Even here, one may point to tax subsidies as potential optimal supplements for the public good associated with the externality aspects of not-for-profit firms. In addition, if uninformed consumers do not realize that they are uninformed, things may go awry. Finally, if firms which can somehow pay out profits easily can take on the not-for-profit form, there can be a "lemons" problem in which firms with true not-for-profit motivation are driven out by cleverly disguised imposters. Whether these impediments are significant remains to be seen, but I would not guess that they are.

In my view, the major message from theoretical or empirical work on not-for-profit health care firms is that such ownership differences turn out to be much less important than they might seem. This is especially so where entry is open, but even under imperfect competition nominal ownership structure seems to matter much less than fundamental economic incentives, particularly when the ownership structure itself is chosen in response to those incentives.

REFERENCES

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, December 1963, *53*, 941–73.

Danzon, Patricia M., "Hospital 'Profits': The Effect of Reimbursement Policies," *Journal of Health Economics*, May 1982, *1*, 29–52.

Fama, Eugene F. and Jensen, Michael C., "Agency Problems and Residual Claims," *Journal of Law and Economics*, July 1983, *26*, 327–50.

Hansmann, Henry B., "The Role of Nonprofit Enterprise," *Yale Law Journal*, April 1980, *89*, 835–901.

Newhouse, Joseph P., "Toward a Theory of Nonprofit Institutions," *American Economic Review*, March 1970, *60*, 64–74.

Relman, Arnold S., "The New Medical-Industrial Complex," *New England Journal of Medicine*, October 23, 1980, *303*, 963–69.

Sloan, Frank A. and Vraciu, Robert A., "Investor-Owned and Not-for-Profit Hospitals," *Health Affairs*, Spring 1983, *2*, 25–37.

Watt, J. Michael et al., "The Comparative Economic Performance of Investor-Owned Chain and Not-for-Profit Hospitals," *New England Journal of Medicine*, January 9, 1986, *314*, 89–96.

Weisbrod, Burton A., *The Voluntary Non-Profit Sector*. Lexington: D.C. Health, 1977.

______ **and Schlesinger, Mark**, "Public, Private, Nonprofit Ownership and the Response to Asymmetric Information," in S. Rose-Ackerman, ed., *The Economics of Nonprofit Institutions*, New York; Oxford: Oxford University Press, 1985, ch. 7.

Williamson, Oliver, *Market and Hierarchies*, New York: Free Press, 1975.

Valuing Health—A "Priceless" Commodity

By Victor R. Fuchs and Richard Zeckhauser*

Price is a fundamental variable in economics. What role, then, can economics play in the analysis of a commodity, like health, that is said to be priceless? A central concern of health economics research has been whether health is a commodity like any other. Our answer to this question is an unwavering "yes and no." Health *is* a commodity: it enters into utility; its supply is not unlimited, but can be increased through the use of scarce resources (broadly defined); the amount that people demand varies inversely with its price. If societal arrangements for protecting, producing, and enhancing health (including the provision of medical care) ignore this aspect of health—if it is assumed to be a truly "priceless" commodity—these arrangements are likely to be inefficient and unsatisfactory.

In many respects, however, health is *un*like other commodities. These differences have profoundly shaped our institutional arrangements—regulatory, financial, organizational—in both the positive and the normative sense. Health is difficult to trade interpersonally. Initial endowments are extremely important. Whereas most commodities are produced by specialists and then sold to the general public, an individual's health status is largely self-produced, being strongly affected by his or her consumption of other commodities. Together, these conditions imply that individuals will not all value health equally at the margin.

Significant externalities of the conventional sort (for example, communicable diseases) afflict health; significant interdependent utilities affect its valuation. The symbolic aspects of health are important. Many people believe health should not be allocated like other commodities (i.e., on the basis of ability to pay). This sentiment may be expressed more often than it is acted on, but even its ability to command lip service is revealing. For example, Joseph Califano—a recent and enthusiastic convert to the competitive market approach in *America's Health Care Revolution* (1986)—tries to soften his message by saying no one should be denied "needed" care.

In this paper we comment on health in relation to such standard economic topics as wealth, time preference, risk aversion, and utility.

I. Health and Wealth

Except in extreme cases it is difficult to sort out the precise nature of the relationship between health and wealth. If wealth is very low, health suffers, sometimes to the point of death. Likewise, a person in very poor health may have no capacity to create wealth. Random shocks to health and wealth are usually positively correlated. Within the characteristic range of health and wealth in developed countries, however, the direction and strength of causal connections are more difficult to ascertain.

Initial endowments of health and wealth are probably positively correlated. Children inherit genes that affect health and mental ability, and they also inherit wealth. Parental efforts during childhood are probably also positively correlated. For example, children who receive above-average investment in health probably also receive more investment in education.

Health and wealth also depend on the choices individuals make as adolescents and adults. Some factors, such as rate of time preference, will lead to a positive correlation, but others, such as a tradeoff between hazardous work and wages, will lead to a negative relationship.

*Stanford University, Stanford, CA 94305, and Harvard University Kennedy School of Government, Cambridge, MA 02138, respectively. We gratefully acknowledge financial support during this work to Fuchs from the Robert Wood Johnson Foundation and to Zeckhauser from the Business and Government Center at Harvard University.

Health seems to be a normal good in the sense that an increase in wealth leads to an increase in the demand for health. However, the wealth elasticity of demand for health may be less than that of some other commodities, some of which (for example, mountain climbing) may have negative effects on health. In poor countries (unlike the United States), cigarette smoking has a fairly high positive wealth elasticity.

Empirical studies, both cross-section and time-series, suggest that the elasticity of mortality with respect to income (a weighted average of the cause-specific elasticities) approaches zero as income rises. This is to be expected because an increase in income tends to reduce the weight of those causes of death that have large elasticities. Consider infant death rates from congenital anomalies (small elasticity) versus deaths from infectious diseases (large elasticity). In poor societies most infant deaths are the result of infectious diseases, and the overall elasticity is large. As societies become wealthier, a much smaller proportion of infant deaths is due to infectious diseases, and the overall elasticity is small.

When health and wealth are distributed unequally, the question of redistribution poses many complex problems. Suppose that health and wealth depend only on initial endowments, so that incentive effects can be ignored. In this hypothetical situation, many economists would argue for redistribution of wealth toward greater equality—but is there a comparable case for redistributing health? If so, how should it be done? In the real world the choice of appropriate redistributive policies is even more complex; inequality in health is partly the result of initial endowments and partly the result of individual behavior and random shocks.

II. Time Preference

Health valuation discussions should focus on outputs rather than inputs—for example, days without morbidity as opposed to physician visits—and should take consumer preferences as the building block for welfare evaluation. Thus, not just lives lost but some variant of quality-adjusted life years (QALYs) seems appropriate in assessing health outputs. Should QALYs be discounted the way cash flows are discounted in investment analyses? Many ethically concerned observers have said no, lest we mistreat the future, and they have been joined by others who oppose technologies that impose long-lived threats to human survival. Economists have tended to support discounting, prescribing that some account be taken of life cycle income patterns, the changing cost of QALY increments over the lifetime, and the higher expected wealth of future generations.

The appropriate principle is quite simple; individuals and policy planners should discount the *value of life years*. Changes in quality due to income or age are counted, as are changes in life expectancy or the costs of health production. If we fail to discount, or do not take account of changing valuations, we will be taking Pareto-dominated actions. With life years, unlike most commodities, there is virtually no exchange among individuals, which implies that individuals will value life years differently. Policy planners thus cannot merely add up discounted life years for different individuals. Rather they should add up the individuals' valuations. (If distribution is a concern, planners should weight a dollar to a rich and a poor person for life years in the same way as if they were deciding on sponsoring operas or repaving streets. The weighting could be one for one if a tax-transfer system is available. Symbolic effects—discussed below—would be added separately.)

Choices that affect the health of future generations (for example, disposal of nuclear waste, biomedical research) are qualitatively different from redistributing health from our present to our future selves. Should the distinctive nature of health affect our intergenerational bequests? Should we be any more or less generous in leaving an intact ozone layer, thus diminishing the risk of future skin cancers, than in leaving up-to-date factories? If we assume no altruism, then we will always benefit by pushing health risks off to future generations. In a world that is generationally selfish, there will be too little *R&D* and too much long-lived pollution. Capital, by contrast, will not be "selfishly" consumed, leaving nothing for the future, since as time rolls forward it can always be sold to the younger generation.

Suppose altruism induces us to leave something to those who follow. Surely we should leave them the mix of health and other endowments that they most prefer among those that cost us the same amount. Allocations to the future then will represent a Pareto-efficient mix, though we will almost certainly fail to make bequests that the future would like to pay for if it could. (With good-specific paternalism, the principle still applies; the net dollar cost of providing different goods to the future will differ, however.)

Most policy planning discussions assume full altruism—future citizens are given equal weight with present citizens—and discount solely for the time value of money. Given this ethical premise, the value of life years to future generations should be discounted at the time-value-of-money rate. If our descendants will be much wealthier than we, our incentive to transfer wealth to them is diminished. But their greater wealth will confer additional value on a life saved in their generation, increasing our incentive to allocate resources so as to improve our descendants' probability of survival. Similarly, just as we tend to value a life saved at age 30 more than one saved at age 80, we should transfer more life-saving resources from current to future generations the greater is their life expectancy relative to ours.

Assessing the present value of life years is just one area where our thinking about discounting has been confused by intergenerational considerations. Economists' discussion of the social rate of discount often founders on the same issue. If it turns out that we collectively value the future more than we express in private actions—for example, I care more about your child's welfare than about yours, but you have no way to charge me for the externalities of your intergenerational transfers—we should adjust our valuations of future benefits upward, not our discount rate downward. Self-respecting economists should not adjust discount rates for externalities stretching to the future or use different rates because it is health that is being valued.

III. Risk and Risk Aversion

Even physicians and epidemiologists have a hard time specifying the production function for health. The commodity is produced probabilistically, substantially through the lifestyle choices of individuals. The relevant probabilities are often small and hard to judge. If A is inefficient in the production of health—say, eating too much meat relative to fish though he has no strong preference between the two—B cannot poach on his market; this suggests the potential for X-inefficiency. The costs of producing additional life years vary significantly, between old and young, as well as between exercise freaks and sedentary types. Moreover, individuals value life years differently because of differences in income, attitudes, and age. In sum, even if there were no income inequalities, individuals would have dramatically different values for the same health output.

Some health valuations merely revolve around probabilities of health loss. From von Neumann-Morgenstern utility, we learn we should be linear with gains in survival probability at any particular time (assuming there is no endowment to be spread across our survival states). In a two-period model, a 1 percent increment in survival probability in the first year is worth a 2 percent gain for the second year, leaving aside time preference and age-related changes in the quality of life. Experiments with hypothetical questions about health suggest that individuals are risk averse with respect to gains but risk preferring with respect to losses, much as Daniel Kahnemann and Amos Tversky have shown for wealth in their work on prospect theory. This suggests how much behavior can depend on mental frame of reference. The point from which an individual perceives gains and losses to occur can influence his choices.

Though life may be priceless—in the sense that there is no monetary amount that an individual would accept to give up his life and (even if there were) society would prohibit its sale—probabilistic risks to life in

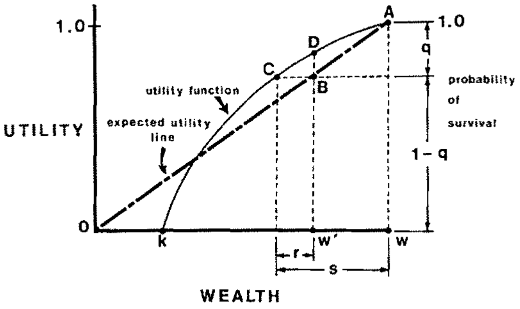


FIGURE 1

exchange for dollars (or other benefits) may be acceptable. Economists have presented this argument in a variety of regulatory arenas, notably those relating to the environment and workplace. In common parlance, an individual who will not take health risks for money is called risk averse. How does this risk aversion relate to the economist's risk aversion, a hesitancy to take gambles on money (i.e., along a single dimension)? The more risk averse a person is with respect to wealth, the more he will pay to boost his probability of survival.

Consider an individual who will either die immediately with probability q, or survive for a single period with probability $1-q$. He will have wealth w should he survive, whereupon he receives utility $u(w)$ at A, whose height is scaled to be 1.0 (Figure 1). Death is assigned a utility of 0. Notice that a wealth of k, perhaps the subsistence wage, is no better than death. The individual's expected utility, as indicated along the heavy dashed line for possible probabilities of survival, is $u(w,q)=(1-q)U(w)+qU(\text{death})=(1-q)U(w)$, as indicated by the height at point B.

What amount of money would the individual sacrifice to avoid the q risk of death? Assume first that his wealth is independent of his survival probability. Then he will sacrifice up to s, where $u(w-s,0)=u(w,q)$, as shown on the diagram by the equal heights for B and C.

On the other hand, if the individual's wealth is an endowment that must be spread in an actuarially fair manner over his survival states, then his income would fall to w', where $(1)w'=(1-q)w$, if he purchased survival probability q, placing him at D. He would then sacrifice wealth of only r to rid himself of death probability q. (The value of r could be negative if he were relatively risk neutral.)

In either case, the amount that an individual would sacrifice to reduce risks to life relates positively to the height of the utility function at current wealth relative to its slope at that point. Holding constant the zero utility wealth, k, greater risk aversion (at all wealths) leads to greater valuation, which accords with anecdotal experience.

To see this intuitively, consider the fixed-wealth case, where willingness to pay to avoid a risk is proportional to height/slope. Scale the utility functions of two individuals with different risk aversion to reach height 1.0 at w. A q risk of death cuts the expected utility of each by the vertical distance q. Willingness to pay to eliminate this risk is the amount of money taken from present wealth that entails the same q reduction in utility. This amount will be greater for the more risk-averse individual, whose utility function is flatter near present wealth.

IV. Policy Issues

The government's role in health is pervasive. Government tells consumers what goods and services they can and cannot buy; it tells workers what jobs and working conditions they can and cannot accept; it spends hundreds of billions of dollars each year subsidizing the demand for health care and regulates every aspect of its delivery. Some of these interventions can undoubtedly be explained by economies of scale in the provision of information or a desire to reduce transaction costs, but most cannot. The demand for extraordinarily expensive health care is highly unpredictable, not unlike the demand for firefighting services. To achieve efficient risk spreading, we can collectivize health care and provide it for free (as Britain has done) or employ insurance, possibly government subsidized (the approach of the United States).

Quite often we are dealing with a situation where A thinks that B is not choosing an optimal level of health and therefore favors legislation that will raise *B*'s health level, even at some cost to *A*. Why does *A* feel this way? One possibility is pure paternalism. *A* thinks he is a better judge of what maximizes *B*'s utility than *B* is, and *B*'s utility matters to *A*. Alternatively, *A* may feel a responsibility to ensure that *B* acts rationally, regardless of any effects on *A*, or to protect *B*-future who is being inadequately considered by *B*-present. Or, sensing that he may find himself in *B*'s shoes some day, *A* may vote for some intervention, say mandatory health insurance, as a form of self-control. Finally, it may not be *B*'s utility in general that matters to *A*, but specifically *B*'s health. *A* is more troubled by *B*'s poor health than by his empty wallet. This could be an atavistic reaction left over from a time when most health problems were communicable. Or perhaps *A* wants to cover up the "raw edges" of inequality in wealth. Thus he is unwilling to let poor people work at hazardous jobs, sell their organs, or otherwise barter health for income.

One popular explanation for the attempts to regulate health behaviors is that *A* does not want to bear the expense of *B*'s health-damaging behavior through collectively financed third-party payments for health care. But *A* also shares in the costs of providing *B*'s retirement income, and thus stands to benefit financially from *B*'s lower life expectancy. On balance, the externalized financial consequences of *B*'s cigarette smoking, including the excise taxes he pays and his lower-than-average expected Social Security benefits received, may more than offset his higher-than-average expected medical costs. If eliminating financial externalities were the key, perhaps society would take more aggressive actions for seatbelts, which disproportionately confer health benefits on individuals entering prime earning years, and less on cigarettes and their mostly older victims. In any case, many ardent "interventionists" seem completely unaware of the financial externalities argument, so we must look elsewhere for an explanation of their preferences.

The symbolic role of health and health care may play a role. Many of the most fundamental beliefs of our society are wrapped up in the valuation of lives and health. Although society cannot avoid making choices that trade health for other resources, the implicit or shadow prices are rarely mentioned; the legitimacy and popular acceptance of the choice process is often as important as the outcome it generates. Because of the special nature of health production, economists cannot make their customary contribution of pointing out how fewer resources would be required to produce the same commodity elsewhere. Better guardrails for *B* in no way compensate for dirtier air or less health care for *A*, however great the net gain in health, and *A* himself is not saving any resources.

That life is priceless need not imply that we will spare no expense to save a life or cure a disease. Yet that myth persists and gives us comfort. When the myth confronts us head on, as in a decision to support renal dialysis for all who need it, we are likely to flinch and provide. Given our need for myths, many mechanisms of cost containment must work in the shadows. If hospital beds or new technologies that are unlikely to be cost effective are put in place, myth maintenance tends to overpower cost containment. Distributional equity also plays a significant role. If a poor man lacks a VCR or even a radio, we are not overcome by guilt. But let him suffer from a lack of health, and all the injustices of our society come into stark relief. It is no surprise that government expenditures on health care for the poor exceed cash transfers to them.

V. Conclusion

We cannot produce life years directly or buy them on any market. Goods that cannot be produced or sold often become priceless through a natural process that leads to interdependent utility considerations, great concerns for equity, probing discussions of obligations to the future, and assertions that these goods shall not be allocated through the market. Many priceless commodities are collectively owned, or at least we act as though we have a collective interest in them. Thus, for example, governments rarely sell their natural wonders or peddle outlying territories to other countries; they prevent the export of rare antiques (the national heri-

Health Economics and Econometrics

By JOSEPH P. NEWHOUSE*

It seems worthwhile to pause occasionally and take stock of the balance of trade between economics and health economics. Martin Feldstein (1974) did this fourteen years ago at the ASSA meetings, and sufficient time has lapsed to bring up the issue again. Here, however, I undertake only a fragment of the task, confining the discussion to some remarks about econometrics and health economics.

If one were writing about economic theory and health economics, one might focus on the importance for health economics of both uncertainty and the physician's dual role as supplier and patient agent. In discussing econometrics and health economics, however, other features of health economics are relevant:

1) It is an applied field with much policy interest. Over 10 percent of GNP is spent in health care, and public programs account for about 40 percent of American expenditure on personal health care (the percentage is even higher in most other countries). The policy interest has at least two implications: (a) The reward for robust, accurate results is considerable, because quantitative results are used in making choices by both public and private decision makers.[1] Put another way, point estimates as well as hypothesis tests are important. (b) The audience for the health economist will be broader than other economists.[2]

2) One of its central concerns is health care expenditures, whose distribution across individuals is very skewed. In any year the highest spending 1 percent of individuals accounts for more than a quarter of the expenditure and the highest spending 5 percent accounts for more than half.[3]

These two features taken together create some tension. The skewness of expenditures obviously poses problems in obtaining robust results. Moreover, simple methods that are accessible to a broad audience may not be very robust.

I. Imports

I begin a balance of trade accounting with imports, which by definition are numerous in an applied field such as health economics. Simply writing down a list of imports, however, is more tedious than enlightening. Rather than take this tack, I concentrate on the waxing and waning in the volume of trade of one prominent import, relatively complete econometric models of the health care sector.

I focus on econometric models because they illustrate the importance of robustness

*The RAND Corporation, 1700 Main Street, Santa Monica, CA 90406. Support for this work was provided by the Rand Corporation and is gratefully acknowledged. A more completely referenced and annotated version of this paper is available from the RAND Corporation. I am grateful to Victor Fuchs, Emmett Keeler, and Willard Manning for comments on an earlier draft. They do not necessarily agree with these views.

[1] For example, outlier days under Medicare are paid at 60 percent of average cost. This value reflects an earlier literature on estimates of marginal cost as a fraction of average cost. There is, however, considerable question as to whether this is too low a figure (Bernard Friedman and Mark Pauly, 1981).

[2] Of course, the audience for many applied fields is broader than simply economists. These comments about health economics, for example, surely apply to macroeconomic forecasting as well.

[3] These results come from the Rand Health Insurance Experiment and exclude spending by the elderly; similar skewness, however, appears to hold for the elderly as well. Victor Fuchs points out that skewness would fall if a time period longer than a year were considered. This is an important observation, but few empirical studies consider periods longer than a year. I suspect longitudinal studies of health expenditures would yield important results, just as the *Panel Survey of Income Dynamics* and the *National Longitudinal Survey* yielded important results in labor economics. Such studies would appear to be a promising research target.

and because of their prominence in Feldstein's earlier survey. Indeed, Feldstein suggests that the refinement of such models can be used as a benchmark to measure progress in health economics. For example, he says:

> Although a consensus has not yet emerged on the correct general specification of a model of the health care sector, a framework for research has been defined.... Although most studies have dealt with only single aspects of the health care system, the "invisible hand" that guides researchers to fill existing lacunae is collecting the pieces that will comprise an econometric model of the entire sector. This process is still far from complete. [pp. 378–79]

It seems doubtful that Feldstein or many others would write in such a tone today, or, if they did, they would give more emphasis to the last sentence. Judging from submissions to the *Journal of Health Economics* and a nonsystematic search of citation counts, few if any economists are currently much interested in an econometric model of the entire health care sector. Why is this?

The purpose of most econometric models was not to advance theory, but rather to quantify relationships and to assist in the policy process. Success in those aims required some degree of predictive accuracy. Unfortunately, the track record of many models was, in retrospect, poor.

A model of dental manpower estimated in the late 1970's provides an example (Nina Mocniak 1981). The model suggested that we were not training too many dentists, something that its creators acknowledged was even then contrary to prevailing opinion. Yet subsequent market behavior supported prevailing opinion; the applicant/enrollee ratio at dental schools fell from 2.5 in 1976 to 1.3 in 1985, while the number of first-year dental students fell from 5,900 to 4,800 over the same period (American Dental Association, 1986).

I am not trying to pick on this particular model (many other examples could be found), but I do note that the model's estimated price elasticity of demand at the mean was −4 (!), which, when combined with the projected spread of dental insurance (a largely accurate prediction), buoyed the predicted demand for dental services and thereby the demand for dentists. One might have thought a price elasticity of −4 would, on its face, have aroused some suspicions about specification or the influence of certain observations, but robustness was not so much in fashion then. If it had been, many (most?) econometric modeling efforts may well not have been undertaken.

II. Transshipment

Because health economics is an applied field, most trade flows are one way; true exports are relatively rare. More common than exports is encouragement or amplification that health economists can give to certain trends that arise elsewhere in economics and econometrics. In this section I focus on two developments in econometrics that health economics could transship, one hopes adding some value in the process.

A. *Robustness and Specification*

As noted above, robustness is important to many health economics studies. Hence, the current emphasis in econometrics on robustness and specification is of obvious importance to health economics. Under the general heading of robustness and specification I group issues of specification tests, robust estimators, and replication.

Specification Tests.[4] The usefulness of specification tests should be clear, but is emphasized by two findings that emerged from the split-sample analysis used in the Health Insurance Experiment (Naihua Duan et al., 1983). First, results can be quite sensitive to the choice of transformation on the left-hand side. For example, one can calculate expenditure by experimental insurance plan using either raw expenditure or the logarithm of expenditure.[5] In part because

[4] See Adrian Pagan (1984) for a discussion of specification tests.

[5] A constant must be added to avoid taking the logarithm of zero. The constant $5, approximately 1 percent of the mean, minimizes the skewness and has been used to derive the values in Table 1.

TABLE 1—INDICES OF RESPONSIVENESS OF MEDICAL CARE USE TO INSURANCE PLAN, TWO ALTERNATIVE DEPENDENT VARIABLES[a]

	First Site Year		First Nine Site Years	
Plan	Raw $	Log ($+5)	Raw $	Log ($+5)
Free Care	100	100	100	100
Coinsurance				
25 percent	109	66	84	72
	(1.36)	(2.37)	(1.38)	
95 percent	57	43	61	45
	(2.99)	(4.30)	(4.51)	
N	841	841	6528	6528

Source: Derived from Duan et al., 1983, Appendix E. Values are unweighted average of nine site years, using ANOCOVA and 1 part model. I have not calculated standard errors for the fourth column because of the difficulty of correcting for intertemporal correlation, but the same pattern is found in each site year, so *t*-statistics would be quite significant.

[a]Free Care = 100; *t*-statistics on contrast with free plan in parentheses.

of the skewness of expenditure, these two specifications of the dependent variable yield substantially different estimates of responsiveness to plan, although the same data and the same right-hand side variables (specified in the same way) are used in both equations (Table 1).[6] Moreover, the standard errors around the estimates using the logarithmic transformation were relatively much smaller than those using the raw dollars, making those estimates superficially more attractive to report. Ultimately the estimates using the raw dollars proved to be more accurate (using the split sample technique).

Second, analysis of variance (i.e., mean expenditure by insurance plan) had lower mean square error in the forecast sample than analysis of covariance, using such simple demographic covariates such as age, sex, race, self-perceived health status, and education of head. Put another way, including these common covariates made one worse off in predicting expenditure. This result, which might be incomprehensible to a person who had simply taken an introductory econometrics course, obviously did not happen because these variables are unrelated to the dependent variable. Rather, the problem stemmed from the skewed distribution of expenditure and overfitting. That is, the covariates tended to fit extreme observations in the estimation sample, and hence did not fit the forecast sample very well.

Neither of these findings would have been uncovered using the standard significance tests of the coefficients to test specification. Indeed, in both cases the standard tests pointed the other way. The *t*-statistics are more impressive for the logarithmic specification and, of course, the standard demographic variables, when included, are quite "significant." Hence, both findings underscore the importance of more omnibus specification tests than the usual *t*-statistics.

Robust Estimators. Robust estimators are similar to transformations as a technique for analyzing skewed data because both effectively downweight extreme observations.[7] Although extreme observations can be very informative, the skewness of health expenditure data poses a serious danger of overfitting using standard least squares methods.

Replication. Replication is a time-honored method for enhancing confidence in both qualitative and quantitative findings. One might suppose that contact with those in medicine and their traditions would make health economists particularly sensitive to those encouraging replication (see William DeWald et al., 1986), but I doubt that the past record of health economists in this regard is any better than other applied fields of economics.[8] It is interesting to speculate why there may be more replication in medical research than in health economics, assuming for the sake of argument that there is. Possibly in medical research it is easier

[6] The economic explanation of the difference is that the log transformation downweights response in the right tail, but the plan response was less in the right tail because of the feature that the coinsurance rate fell to zero after the family spent $1000 out of pocket.

[7] See William Krasker (1986) for a recent application.

[8] For a useful attempt by health economists to replicate and extend, see Jerry Cromwell and Janet Mitchell (1986).

for researcher 2 to generate researcher 1's data without researcher 1's cooperation (by rerunning the experiment) than in health economics. Of course, this merely raises the question of why data sharing of the kind Dewald et al. discuss is not more widespread in economics.

B. *Experimentation*

A second development in economics and econometrics that should assist in obtaining more robust results is experimentation. And, indeed, health economics has used experiments, most notably the Health Insurance Experiment (see my 1974 paper; Willard Manning et al., forthcoming),[9] as well as smaller scale field experiments (David Greenberg and Philip Robins 1986). Laboratory experiments as well are useful in breaking new ground. For example, Charles Plott and Louis Wilde's (1982) experiments on demand with asymmetric information seem like a promising avenue with which to investigate the issue of supplier-induced demand.

Controlled experiments, of both the field and laboratory variety, have several well-known advantages over observational studies. In particular, by keeping the treatment exogenous and (in general) uncorrelated with other explanatory variables, they yield more precise results than observational studies for a given (finite) sample size.[10] The usually cited disadvantage is cost, but this applies only to field experiments, and even in that case the additional cost of an experiment relative to a prospective observational study is likely to be relatively small.

III. Exports

I close with two examples of tools developed to solve problems in health economics that appear to have wider applicability. Both came from work done for the Health Insurance Experiment.

Retransformation: Long before anyone heard of health economics, econometricians were using logarithmic and other transformations. Generally the issue of retransforming to the original scale did not arise. For example, if a Cobb-Douglas production function was estimated with the log of quantity on the left-hand side, econometricians rarely used the equation to predict quantity in raw units. Usually the emphasis was on the elasticities, and one did not need to retransform to estimate the elasticities.

If, however, one transforms to minimize the effect of skewness (for example, takes the logarithm of dollars expended), but still wants to know the response surface in raw units (for example, in dollars rather than log dollars), a retransformation is necessary. In this example, if the error term is lognormally distributed, one can exponentiate the sum of the predicted log and half the variance. Estimates using this formula, however, are sensitive to seemingly minor departures from lognormality in the right tail.

Duan developed a nonparametric estimator for retransformation (the "smearing" estimator) that is both easy to use and does not require parametric assumptions on the error term.[11] It requires that the error term be independent of the explanatory variables in both expectation and higher moments; the expectation condition, however, is required for consistent estimation anyway. Moreover, when the error is lognormal, the estimator loses little to the parametric estimator. The smearing estimator is now beginning to be used (Robert Ohsfeldt and Steven Culler, 1986).

The Allocation of Subjects to Treatments in an Experiment and the Choice of Units to Sample in an Observational Study. The classical methods for allocating subjects to experimental treatments are simple randomization or randomization within strata and blocking designs. These techniques, however, have two drawbacks: 1) for computational reasons one

[9] In light of the prior social experiments in income maintenance, labor economists might say health economics imported field experiments.

[10] Contrast the analysis of the Health Insurance Experiment data (Manning et al.) with myself and Charles Phelps (1976).

[11] In the case of the logarithmic transform, the retransformation factor is the average of the exponentiated residuals.

can stratify or block on relatively few dimensions; and 2) continuous variables such as income or age must be grouped into discrete intervals; hence, within interval variation is lost.

The Finite Selection Model (Carl Morris, 1979) begins with a (finite) list of persons to be allocated to treatments; the intent is to allocate the list in such a way that the distribution of characteristics on each treatment is similar to the distribution on every other treatment. The characteristics can be treated as continuous variables (if they are continuous), and they can be weighted in importance (i.e., the distribution of some characteristics could be made more similar than other characteristics). In the case of the Health Insurance Experiment, the design obtained using the Finite Selection Model yielded standard errors that were on average about 25 percent less than those that would have been obtained from simple random allocation.

The model can also be used in observational studies to choose which units to sample. Used in this fashion, it chooses a sample that most closely represents the distribution of the population along dimensions specified by the analyst; for example, which 50 metropolitan areas most closely represent the universe of metropolitan areas. Thus, the model affords protection against the possibility that simple random sampling yields an unrepresentative sample through bad luck.

REFERENCES

Cromwell, Jerry and Mitchell, Janet B., "Physician Induced Demand for Surgery," *Journal of Health Economics*, December 1986, *5*.

DeWald, William G., Thursby, Jerry G. and Anderson, Richard G., "Replication in Empirical Economics: *The Journal of Money, Credit, and Banking Project*," *American Economic Review*, September 1986, *76*, 587–603.

Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, September 1983, *78*, 605–10.

______ **et al.**, "A Comparison of Alternative Models of the Demand for Medical Care," *Journal of Business and Economic Statistics*, April 1983, *1*, 115–26.

Feldstein, Martin S., "Econometric Studies of Health Economics," in M. Intriligator and D. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland, 1974.

Friedman, Bernard and Pauly, Mark V., "Cost Functions for a Service Firm with Variable Quality and Stochastic Demand," *Review of Economics and Statistics*, November 1981, *63*, 610–24.

Greenberg, David H. and Robins, Philip K., "The Changing Role of Social Experiments in Policy Analysis," *Journal of Policy Analysis and Management*, Winter 1986, *5*, 340–62.

Krasker, William S., "Two-Stage Bounded-Influence Estimators for Simultaneous-Equations Models," *Journal of Business and Economic Statistics*, October 1986, *4*, 437–44.

Manning, Willard G. et al., "Health Insurance and the Demand for Medical Care: Results from a Randomized Experiment," *American Economic Review*, forthcoming June 1987.

Mocniak, Nina, "Aggregate Supplies and Demand of Dental Services," in L. J. Brown and J. E. Winslow, eds., *Proceedings of a Conference on Modeling Techniques and Applications in Dentistry*, DHHS Publ. No. (HRA) 81–8, Washington: US GPO, 1981.

Morris, Carl, "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics*, September 1979, *11*, 43–61

Newhouse, Joseph P., "A Design for a Health Insurance Experiment," *Inquiry*, March 1974, *11*, 5–27.

______ **and Phelps, Charles E.**, "New Estimates of Price and Income Elasticities for Medical Care Services," in Richard N. Rosett, ed., *The Role of Health Insurance in the Health Services Sector*, Universities-National Bureau Conference Series, No. 27, 1976.

Ohsfeldt, Robert L. and Culler, Steven D., "Differences in Income between Male and Female Physicians," *Journal of Health Economics*, December 1986, *5*.

Pagan, Adrian R., "Model Evaluation by Variable Addition," in D. F. Hendry and Kenneth F. Wallis, eds., *Econometrics and Quantitative Economics*, Oxford: Basil Blackwell, 1984.

Plott, Charles R. and Wilde, Louis L., "Professional Diagnosis vs. Self-Diagnosis: An Experimental Examination of Some Special Features of Markets with Uncertainty," in V. L. Smith, ed., *Research in Experimental Economics*, Vol. 2, Westport: JAI Press, 1982.

American Dental Association, Council on Dental Education, *Dental Education Trend Analysis*, Chicago: Division of Educational Measurement, ADA, 1986 (Suppl. 11 to ADA Annual Report on Dental Education for 1985–86).

STOPPING HIGH INFLATION[†]

The Israeli Stabilization Program, 1985–86

By STANLEY FISCHER*

The success of the Israeli stabilization plan was far from a foregone conclusion to those who urged and implemented it. That which is now commonplace—that heterodox stabilization measures can bring a quick and almost painless end to rapid inflations—was only a theoretical possibility to the creators of the independently conceived Argentinian plan that went into effect in mid-June 1985 and the Israeli plan that went into effect July 1, 1985.[1]

Evaluation of the stabilization program at this stage must be preliminary—and for that reason more valuable. An account written in the midst of uncertainty about the eventual success of the program gives a more accurate picture of the policy environment than an historical perspective written after the dust and the interest have settled.[2]

I. The Buildup to the Stabilization Plan

Israel stands out among the countries that attempted heterodox stabilization plans (Rudiger Dornbusch and Mario Simonsen, 1987) as having had a reasonably short inflationary history. Inflation rose in steps from 2 percent in 1967 to 500 percent in 1984. Indexation, introduced for labor contracts during World War II, and for capital contracts in the early 1950's, was in place for nearly two decades of only moderate inflation.[3] Behind the inflation was a double-digit budget deficit (even on an inflation adjusted basis) and an accommodating monetary policy. (See Table 1.)

The beginning of large-scale U.S. (initially military) aid in 1974, the move to a crawling peg exchange rate in 1975 and "liberalization" in the form of greater access to foreign exchange and to foreign-exchange linked accounts, introduced at the end of 1977, were key changes in the institutional structure of the economy. From 1977 to 1979, the economy made the transition to triple-digit inflation. The inflationary jump appears to have been a slow reaction to the shift out of domestic money that accompanied the liberalization: the growth rate of M_1 actually fell in 1979 as inflation accelerated.

At that time the predominant view was that Israelis, protected by extensive indexation, could live with inflation and that unemployment was at all costs to be avoided. This ruled out orthodox inflation stabilization policy. An attempt was made from 1981 to slow inflation by reducing the rate of devaluation and the rate of increase of controlled prices. The policy led, as it had in Latin America, to a growing overvaluation of the currency and current account deficit. Inflation did not appreciably slow.

The increasing overvaluation of the currency and growing current account deficit made a discrete devaluation inevitable. At the end of 1983 there was a stock market

[†] Peter Garber, Brown University; Guillermo Calvo, University of Pennsylvania.

*Department of Economics, MIT, Cambridge, MA 02139, and Research Associate, NBER. I have benefited from discussions with Michael Bruno and Rudiger Dornbusch. Research support was provided by the National Science Foundation.

[1] The genesis and structure of the Israeli stabilization plan are described by Michael Bruno (1986a, b). Rudiger Dornbusch's and my paper (1986), written in the fall of 1985, presents a preliminary account.

[2] Accounts of the classic stabilizations make success appear more inevitable than it must have been. One gets as much insight into the precariousness of the Austrian stabilization of 1922 by reading J. van Walre de Bordes (1924) hedging his bets on its success as by reading his chronology.

[3] The inflationary history of Israel is reviewed in my papers (1984; 1985), and in Bruno's and my paper (1986).

TABLE 1—PRESTABILIZATION ECONOMIC DATA

Period	Inflation	GNP Growth	Budget Def./ GNP[b]	M_1 Growth
1960:1–1973:3	7.6	8.5	2.7	17.5
1973:3–1977:2	36.0	2.6	17.6	26.5
1977:2–1979:4	71.0	3.0	17.2	37.0
1979:4–1983:3	123.3	1.6	14.1	97.1
1983:3–1985:2[a]	398.6	5.3	17.0	310.4

Source: Bruno and myself (1986), updated from *Monthly Bulletin of Statistics*, Central Statistical Office, Israel, and *Bank of Israel Report*, 1985.
Note: All growth rates are at an annual rate.
[a] Growth rate of GNP for this period sensitive to choice of initial quarter.
[b] Shown in percent.

crisis as the result of a portfolio shift into foreign-currency-linked assets, a massive devaluation, and the resignation of the Finance Minister the same day he proposed dollarizing the economy. The fury of the nationalist reaction made it clear that possession of a national currency is, at least for Israelis, one of the attributes of sovereignty.

Devaluation at the end of 1983 kicked the inflation rate up to a new plateau of nearly 500 percent per annum. Despite an improving current account caused by the devaluation and the U.S. recovery, a capital outflow continued through 1984. The steady loss of reserves led to uncertainty about Israel's ability to service its external debt, which amounted to 70 percent of GNP, and to rumors that commercial banks were reluctant to roll it over.

It was confidently expected before the July 1984 election that the new government would attack both the inflation and the balance of payments problems, and that the attack would include a significant cut in the budget deficit. Despite the inflation, and the war in Lebanon, there was no decisive winner of the election. It was September before the new national unity government was formed. All was ready for a fully expected comprehensive stabilization program.

II. The Stabilization Program

Instead nothing much happened. The Prime Minister, from the Labor Party, was not predisposed to tough budgetary measures and the possibility of unemployment, and hoped to reduce inflation and solve the balance-of-payments problem through a social compact among labor, the employers and the government, and increased U.S. aid. Inflation for the three months, August to November 1984, was at an annual rate of 950 percent.

In November the government brokered a package deal in which labor agreed to a wage freeze and employers agreed to hold the line on prices. The inflation rate dropped to 4.5 percent per month over the next two months, but the package could not hold. The budget deficit had not seriously been cut and devaluations were continuing. Inflation resumed, averaging 350 percent for the first six months of 1985. Reserve outflows were continuing at the rate of $200 million per month; by June reserves had fallen to $2 billion, compared with a normal $3 billion (GNP is about $25 billion). The black market exchange premium was above 30 percent. Although the U.S. government had indicated it would provide Israel with supplementary emergency aid of $1.5 billion over the next two years,[4] it showed reluctance to be forthcoming until Israel took decisive action.

Without the active involvement of the Prime Minister, Israel's coalition government, in which ministries and their budgets are the prizes in the party political game, cannot make decisions on major economic issues. In June, the Prime Minister committed himself to support a comprehensive stabilization program that would include a substantial budget cut and devaluation. A small team of economists from inside and outside the government was appointed to work out the details of the program, with the

[4] Regular aid in fiscal year 1985 was $3 billion, $1.8 billion in military and $1.2 billion in economic aid, all in the form of grants. Interest payments on debts incurred after the Yom Kippur War were approaching $1 billion.

participation of the Prime Minister and Finance Minister. The program was pushed through the cabinet in a close to 24-hour session ending July 1.

The key points of the program were: Cuts in the government budget, mainly in subsidies, amounting to 7 percent of GNP; Devaluation of 19 percent and fixing of the exchange rate to the dollar; Suspension of wage contracts (including indexation) and the granting of a general wage increase of 14 percent, pending negotiations between the employers and the Histadrut (the national trade union); A price freeze; Israelis would no longer be able to hold dollar-linked short-term deposits; The Bank of Israel would conduct monetary policy with the exchange rate as its main nominal target and credit as a nominal indicator.

The program was inevitably the result of a series of compromises between the desire for measures that would decisively solve the inflation, budget, and balance of payments problems, and the need for political support. Budget cuts were taken as far as political opposition allowed; in particular, it was difficult to make significant cuts in government purchases of goods and services, without reducing the defense budget which had already been cut in 1984. The devaluation was limited by an evaluation of how far real income could be reduced. A capital levy was rejected, partly because a law forbidding taxation of savings had been passed before the July 1984 election and it was feared that the Knesset would vote down any attempt to bring new legislation. The government also feared that its future ability to borrow would be impaired by any such measure.

Most important of all, and the rationale for the program, was the desire and political necessity of avoiding unemployment. The analysis was that a new low inflation equilibrium was attainable with significant budget cuts,[5] and that a coordinated program would move the exchange rate, wages, and prices to new equilibrium levels without the protracted unemployment that would be needed to force the rate of nominal wage increase down without controls. That is the new heterodoxy.

The cutting of subsidies and devaluation raised the price level 27.5 percent in the first month of the program. Labor objected to the use of emergency legislation to suspend labor contracts, and some strikes took place—but without much public support.[6] A new wage agreement was reached in the middle of July, making level adjustments in the wage, reinstating COLA agreements at the end of the year, and providing nominal wage increases amounting to 12 percent over three months starting with the December 1985 wage. Essentially the government was given six months of stable nominal wages (after the adjustments). A reasonable evaluation at the time, and certainly in retrospect, is that the government should have brought more pressure to bear on the wage negotiations, and that the agreement to nominal wage increases even in six months time was a mistake. Even with the unanticipated fall in the price of oil in 1985–86, wages have risen faster than is consistent with maintenance of the exchange rate.

III. The Outcome and Prospects.

Table 2 presents quarterly data since mid-1985. The outcome is impressive. The inflation rate came down to a 1–2 percent per month rate within three months. There were few shortages and most price controls have been lifted.[7] The domestic budget deficit remains low and, with foreign aid, the budget is in surplus. After a short period of crisis early in 1986 when the government showed signs of kicking over the traces, budget discipline has been maintained. Unemployment

[5] Bruno and I (1985), among others, have written on the high inflation trap that is a result of dual equilibria in inflationary economies. As far as I know, neither of us ever argued or believed that Israel could end inflation without significant cuts in the budget deficit.

[6] The Histadrut (the national labor union) had indicated before the plan its willingness to share the burden of stabilization. Necessary atmospherics aside, its behavior after the plan was announced was not uncooperative.

[7] The price controllers were able to take advantage of falling oil prices by reducing the price of fuel as other prices went up when controls were lifted.

TABLE 2—POSTSTABILIZATION DATA

	Inflation	Budget Deficit/ GNP[a]	Unemployment[a]	Real Wage Index	Exchange Rate Index	Credit Growth
1985, to June	11.9	12.5	6.3	113.6	64.4	12.3
1985:III	10.9	5	7.5	95.6	79.8	8.9
1985:IV	2.1	4	6.7	93.5	77.4	1.7
1986:I	0.6	3	7.2	106.4	70.3	2.9
1986:II	2.2	2	7.9	112.0	64.3	2.4
1986:III	1.0	4	6.8	111.1	63.9	2.4

Sources: Monthly Bulletin of Statistics, Central Statistical Office, Israel, and Bank of Israel. Budget deficit data are Bank of Israel estimates. Exchange rate index measures the wage relative to the exchange rate of a currency basket.

Note: All growth rates are for last month of period relative to last month of previous period, at a monthly rate.

[a] Shown in percent.

has risen less than 2 percent. The current account (not shown) has maintained its improvement. The black market exchange premium is zero.

Problems of course remain, in two areas. First, the stabilization is not assured. Most important, a key to maintaining low inflation has been the fixed exchange rate. Although payroll tax reductions have reduced the cost of labor to employers, the currency is clearly becoming overvalued. At some point the exchange rate will have to be adjusted.[8] A way has to be found to make that transition without restarting the inflationary spiral, and to enable Israel to move to world inflation rates.

Second, the economy needs major structural reforms. The ending of inflation was a necessary precondition for dealing with the economy's real problems.

REFERENCES

Bruno, M., (1986a) "Sharp Disinflation Strategy: Israel 1985," *Economic Policy*, April 1986, *2*, 379–402.

______, (1986b) "Israel's Stabilization: The End of the 'Lost Decade'," mimeo., Falk Institute, Hebrew University of Jerusalem, 1986.

______ **and Fischer, S.,** "Expectations and the High Inflation Trap," mimeo., Department of Economics, MIT, 1985.

______ **and** ______, "The Inflationary Process: Shocks and Accommodation," In Yoram Ben Porath, ed., *The Israeli Economy*, Cambridge: Harvard University Press, 1986.

Dornbusch, R. and Fischer, S., "Stopping Hyperinflations Past and Present," *Weltwirtschaftliches Archiv*, 1986, Band 122, Heft 1, 1–47.

______ **and Simonsen, M. H.,** "Inflation Stabilization with Incomes Policies Support," Group of 30, New York, 1987.

Fischer, S., "The Economy of Israel," in *Carnegie-Rochester Conference Series on Public Policy: Monetary and Fiscal Policies and Their Application*, Spring 1984, Vol. 20, 7–52.

______, "Inflation and Indexation: Israel," in John Williamson, ed., *Inflation and Indexation*, Institute for International Economics, 1985.

van Walre de Bordes, J., *The Austrian Crown* (1924), New York: Garland Publishing, 1983.

[8] A 10 percent devaluation took place in January 1987. The Histradut agreed to a 2.7 percent cut in the real wages and payroll taxes were cut, reducing the real cost of labor to employers by 5.4 percent in total.

The Bolivian Hyperinflation and Stabilization

By JEFFREY SACHS*

The inflation in Bolivia during 1984 and 1985 was the most rapid in Latin American history, and one of the highest in world history. During the twelve-month period, August 1984 to August 1985, prices rose by 20,000 percent, and during the final months of the hyperinflation, from May 1985 to August 1985, the inflation surged to an annualized rate of 60,000 percent. As in other hyperinflations, the end came abruptly. A new government came to power in early August 1985, and a comprehensive stabilization program was announced on August 29, 1985. Within ten days the inflation was halted, and prices actually began to fall. Although prices jumped again in December 1985 and January 1986, prices rose by just 9 percent between the weeks of January 20, 1986 and November 3, 1986.

The Bolivian inflation is the only case in thirty-five years of a "true" hyperinflation, applying Phillip Cagan's 1956 classic definition of price increases exceeding 50 percent per month (by this definition, the hyperinflation lasted from April 1984 to September 1985). As such, the Bolivian case provides an important opportunity for examining alternative views of hyperinflation and price stabilization. My discussion of the Bolivian experience is as follows (a more complete account can be found in my 1986 paper, from which this paper draws). Section II describes the chronology of the hyperinflation and stabilization from 1980 to 1986. Section III then examines the rapid end of the hyperinflation in view of competing theories of price stabilization.

I. A Description of the Hyperinflation and Stabilization Periods

The Bolivian inflation may be unique in the annals of hyperinflation as the only case that did not arise in the aftermath of a foreign war, a civil war, or a political revolution (see Forrest Capie, 1986, for a discussion of the background to the hyperinflations of the twentieth century). Rather, the hyperinflation emerged as the result of several less-dramatic shocks hitting the country during a period of intense political instability. Bolivian regimes have been notoriously weak throughout the country's history, but the instability during 1979–85 exceeded the country's norm. A bewildering series of coups, electoral stalemates, and interim governments led to a remarkable turnover of heads of state after 1978.

The instability arose from political and economic disturbances following several years of relative stability and prosperity under the regime of General Hugo Banzer (1971–78). The prosperity of the Banzer era was based to a large extent on Bolivia's favorable terms of trade in the 1970's, heavy foreign borrowing from the international banks, and a political regime supportive of foreign direct investment in the country. When General Banzer left office in 1978, partly under U.S. pressure to restore civilian rule in the country, there ensued a bitter political competition for power, pitting the left against the right and civilian politicians against the military. A worsening international economic environment after 1980, with high interest rates, falling commodities prices, and tight credit, added to the instability.

By the end of 1980, Bolivian access to the private international capital markets had dried up. An emergency rescheduling with the commercial banks was completed in 1981, but it soon fell apart. The World Bank and the IMF also ceased lending. The economic situation deriorated so much that the military returned to the barracks in 1982. Siles Suazo, who had received the plurality of votes in a stalemated election of 1980, was allowed to take power as head of a leftist coalition government in October 1982. This

*Harvard University, Cambridge, MA 02138—Economic Advisor to the Bolivian government. The views expressed here are my own.

government lasted until it was voted out of office in July 1985.

The Siles government inherited an annual inflation rate of approximately 300 percent (October 1982 over October 1981), an inability to borrow on international markets, and an economy declining sharply in real terms (real GNP fell by 6.6 percent in 1982). At the same time, the new government was called upon to satisfy pent-up social and economic demands. The demands on the government were heightened by the fact that the government represented a coalition of forces on the political left that pressed for increases in social spending, public sector employment, and public sector pay, but that lacked the political base to raise tax revenues to fund such spending. The Siles government actually proposed several stabilization programs during 1982–85, but in each case these programs were overturned by public protest, by key constituencies of the government, or by the government's political opposition in the Congress. The inflation tax, unlike other proposed spending or tax measures, did not provoke a specific and well-organized challenge to the government, though in the end the hyperinflation forced early elections, and the ouster of the Siles government.

The precise quantitative links of inflation to foreign borrowing and fiscal policy are difficult to measure because of the incompleteness of Bolivian data. Nonetheless, we may identify three fundamental aspects of the rise of hyperinflation. First,, the cutoff in international lending and the increase in international interest rates in the early 1980's were the main initiating factors in the outbreak of hyperinflation, since the government met the drop in net international financing and higher debt-servicing costs through an increase in the inflation tax. World Bank data on net resource transfers to the Bolivian public sector from medium- and long-term foreign borrowing indicate a dramatic shift in international lending conditions.

With Bolivian GNP approximately equal to $3.6 billion, the shift from net resource transfers *towards* Bolivia of $178 million in 1980 to net transfers *away* of $190.0 million in 1983 signifies a shift of approximately 10 percent of GNP. In the first half of 1985, the Siles Administration ceased virtually all debt service payments to the private creditors, and the Paz government continued that moratorium on debt servicing.

The second important characteristic of the inflation dynamics is that the recourse to seignorage (i.e., the inflation tax) jumped as the net international resource transfer turned negative, with seignorage fluctuating around a new high plateau during the period of the Siles government.

The takeoff in inflation after 1981 followed closely upon the jump in seignorage, as should be expected from almost any monetary theory of hyperinflation. However, during 1982–85 the inflation rate continued to accelerate even though seignorage collection did not rise steadily after its one-time jump. This pattern is familiar from Cagan's original model of hyperinflation. Assuming that expectations of inflation adjust slowly to actual inflation, and that the demand for real money balances is a function of expected inflation, a permanent increase in the seignorage needs of the government (as occurred after 1981) will produce an inflation rate that rises over time while the stock of real money balances declines. If the new seignorage level is sufficiently high, then inflation will rise without bound in the Cagan model.

The third important aspect of the inflation dynamics was that the increasing inflation caused the tax system to collapse, with total central government revenues falling from around 9 percent of GNP in 1981 to about 1.3 percent of GNP in the first half of 1985. Therefore, the constancy of the seignorage tax did not reflect a constant underlying path of real revenues and real expenditures, but rather a path of falling real tax collections matched with a lag by a steady drop in real spending. As described in my 1986 paper, the most important cuts in spending were made in public investment and then in international debt-servicing (as the country went into virtually complete default to private creditors by the end of 1984). Ironically, by the peak of the hyperinflation in mid-1985, debt-servicing payments were almost completely eliminated. The original factor that had started the hyperinflation was gone, but by now the tax system was in a shambles.

The successful stabilization program was carried out by the newly elected center-right government of President Victor Paz Estenssoro in August 1985. The so-called New Economic Policy of the Paz government, unveiled on August 29, 1985, contained an enormously ambitious agenda, that went beyond macroeconomic stabilization to include fiscal reform, trade liberalization, internal price decontrol, and the decentralization or privatization of public enterprises. On the macroeconomic side, the program contained four explicit elements, and an implicit fifth element. The four explicit elements were: 1) a devaluation and subsequent managed float of the exchange rate, and a commitment to full currency convertibility on the current and capital accounts; 2) an immediate reduction in the fiscal deficit through a sharp increase in public sector prices (especially the price of domestic oil), combined with a public sector wage freeze; 3) a tax overhaul proposal (enacted by the Congress eight months later) to broaden the tax base and raise tax revenues; and 4) the signing of an IMF standby arrangement (accomplished in June 1986), to be followed by a Paris Club rescheduling of government debt owed to foreign official creditors. Of these steps, the one with the most important short-run effect was the rise in public sector prices, which raised government revenues immediately by several percent of GNP. The implicit fifth element was the continued complete moratorium on repayments of principal and interest to the commercial bank creditors, despite the strong urgings of the IMF to resume debt servicing.

The policy package had the desired effect of closing the flow budget deficit of the central government. With the combination of higher public sector prices, the virtual halt to all public investment, a tight freeze on public sector wages at very depressed levels, and a moratorium on foreign debt servicing, government revenues jumped above expenditures after the start of the program. The central government did not rely at all on fiscal credit from the central bank during the final months of 1985 or during the year 1986.

II. Interpreting the Sudden End of the Hyperinflation

Despite an enormous devaluation at the beginning of the program, and increases in other key prices, the overall price level stabilized almost immediately. The sudden end of a 60,000 percent inflation seems almost miraculous, except for the fact that virtually every hyperinflation has stopped in that manner. In a justly renowned study, Thomas Sargent (1986) argued that such a dramatic change in price inflation results from a sudden and drastic change in the public's expectations of future government policies, and therefore argued that a convincing "regime change" is the necessary and sufficient condition for rapid disinflation. I suggest, in distinction to Sargent, that the Bolivian experience highlights a different and far simpler explanation of the very rapid end of hyperinflations. By August 1985, the U.S. dollar and not the Bolivian peso was satisfying two of the three classic roles of money: the unit of account and the store of value (though it was not the medium of exchange for most transactions). Prices were set either explicitly or implicitly in dollars, with transactions continuing to take place in peso notes, at prices determined by the dollar prices converted at the *spot* exchange rate. Therefore, by stabilizing the exchange rate, domestic inflation could be made to revert immediately to the U.S. dollar inflation rate!

In my earlier paper, I showed ecometrically that the key therefore to ending the price inflation was to halt the depreciation of the black market exchange rate (which became the same as the official rate once rates were unified through an open auction system). Fortunately, pegging or stabilizing an exchange rate in the short term is far simpler than convincing the public that a true and fundamental "regime change" has occurred. An exchange rate can be temporarily pegged, for example, even if the public knows *certainly* that the peg will break down in the near future, as shown in the literature on rational speculative attacks against currency pegs. In the Bolivian case, the government was able, at least temporarily, to stabilize the exchange rate because it had at least temporarily eliminated its flow budget deficit. The public, however, remained deeply skeptical that the government would be able to maintain the budget austerity beyond a short period (especially in view of the fact that several previous stabilization attempts had failed).

There are two major pieces of evidence in support of the proposition that price stabilization preceded the credibility of the program by several months. First, interest rates on peso-denominated loans remained very high in the months after the stabilization, much higher than comparable dollar-denominated loan rates. This suggests that even though price stabilization proceeded rapidly, and the peso exchange rate stabilized vis-à-vis the dollar, during late 1985 and throughout 1986, the public continued to expect a significant rate of peso depreciation. Indeed, monthly peso interest rates did not fall below 5 percent per month until October 1986, after 8 months of virtual exchange rate and price stability.

The second piece of evidence on the slow return of confidence in the success of the stabilization program comes from the behavior of real money balances. In 1983, real money balances stood at 10.3 billion 1980 pesos. On the eve of stabilization, real balances fell to 3.0 billion 1980 pesos. After price stability was restored in Bolivia, real money balances increased only very gradually in response. As late as June 1986, real money balances still stood at 3.6 billion 1980 pesos, far below the 1983 levels.

The Bolivian authorities were able to maintain the stable exchange rate over the longer run because of the shift from fiscal deficit to fiscal balance, so that for the longer term, Sargent is correct to stress the need for a fundamental change in fiscal policy in order to sustain price stability. In this regard, the Bolivian government instituted a major tax reform in May 1986 that will greatly broaden the revenue base, and it also instituted a very austere budget for 1986. Moreover, it has maintained the principle of no debt-service payments to the commercial banks until it can conclude a fundamental settlement of the debt problem involving some form of debt forgiveness.

The conclusion, then, is that exchange rate stabilization played the preeminent role in the immediate disinflation, while fundamental fiscal policy changes were necessary to maintain the stable exchange rate over time. In other countries, however, such as Chile during 1979–81, the exchange rate has been stabilized but *without* an immediate disinflation. The difference between Bolivia and these other cases seems to lie in nature of hyperinflation. With prices rising at over 50,000 percent per year, all vestiges of long-term pricing in the domestic currency disappear, so that fixing the nominal exchange rate becomes sufficient for price stabilization. For less rapid inflations, staggered price contracts denominated in the domestic currency or backward-looking indexation to the domestic CPI can break the tight link between inflation and the exchange rate. Thus, countries with high inflations but not hyperinflations might be wise to combine exchange rate stabilization with other measures (for example, incomes policies) in order to achieve an immediate return to price stability. Such "heterodox" strategies have recently been employed in Argentina, Brazil, and Israel, in contrast to the "orthodox approach" of the Bolivian authorities. In all cases, however, longer-term stabilization will almost surely require a rectification of more fundamental fiscal imbalances.

REFERENCES

Cagan, Phillip, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956.

Capie, Forrest, "Conditions in Which Very Rapid Inflation has Appeared," in K.

Brunner and A. H. Meltzer, eds., *Carnegie-Rochester Conference Series on Public Policy: The National Bureau Method, International Capital Mobility, and Other Essays*, Vol. 24, Spring 1986.

Sachs, Jeffrey, "The Bolivian Hyperinflation and Stabilization", NBER Discussion Paper Series No. 2073, November 1986.

Sargent, Thomas J., "The Ends of Four Big Inflations," in his *Rational Expectations and Inflation*, New York: Harper & Row, 1986.

The Austral Plan

By DANIEL HEYMANN*

In recent years, several countries have experienced extremely rapid inflations, but without total collapses of the national currencies. Argentina was one of those cases. By mid-1985, the rate of price increase exceeded 30 percent per month; however, prices continued to be set in pesos and there remained a sizeable volume of (short-term) nominal contracts.

The treatment of such inflations poses several questions: What is the desirable speed of disinflation? What mix of instruments can effectively act on prices without causing an excessive fall in real income? How can prices be kept on a stable or moderately rising trend? The Austral Plan, announced in June 1985, combined fiscal measures with a price-wage freeze and a monetary reform, linked to a system for the conversion of debt contracts. The program quite successfully managed the transition to a much lower inflation rate, while real output recovered rapidly after an initial contraction. Still, inflationary forces remained strong: the problem of achieving a sustained stabilization proved quite difficult to solve. This paper briefly describes the program and its effects, with some references to current debates on stabilization policies. (See my 1986 study for full discussion and details.)

I. The Initial Conditions

The Argentine economy went through a drastic real adjustment in the early 1980's. Between 1980 and 1984, the trade balance passed from a deficit of 4 percent of GDP to a surplus of 5 percent, industrial output fluctuated sharply around a declining trend, while investment steadily decreased. Inflation accelerated from 90 percent per year in 1980 to around 700 percent in 1984. This process was marked by successive shocks—large real devaluations, a financial reform that set negative real interest rates in order to reduce business debts, and, later on, a strong wage push. These shocks caused discrete jumps in the inflation rate. The general use of explicit or implicit indexing schemes propagated the effects of the shocks, while aggregate nominal demand sooner or later rose to accommodate the higher inflation.

The public sector deficit rose to more than 15 percent of GDP in 1983. Although in the following year the constitutional government managed to reduce it somewhat, the deficit remained very large. After 1980, the value of interest due abroad rose sharply, partly because the government absorbed much of the private debt. In addition, the "fiscal lag effect" and more widespread tax evasion diminished receipts. With few alternative sources of financing, the Treasury was led to demand massive loans from the central bank.

Relative prices varied much during this period. In addition to the "noise" created by inflation, there were more systematic shifts. By 1984, the real exchange rate was 160 percent higher then four years before,[1] while public sector prices had increased by around 50 percent. In the first part of 1985, the authorities raised these prices further, in preparation for the stabilization program. Real wages oscillated widely, with a sharp minimum in 1982 and a very rapid recovery until late 1984; they fell in the first part of 1985, but were still noticeably higher than at their troughs.

The 1980's inflation was even more rapid than had been usual in Argentina. Money demand dropped to around 2.5 percent of GDP by mid-1985. People used various

*Economic Commission for Latin America, Buenos Aires Office, and University of California, Los Angeles, CA 90024. I alone am responsible for the opinions expressed in this paper.

[1] On a purchasing power measure based on the nominal exchange rate and the CPI inflation differential with the United States.

means to synchronize payments and receipts. In particular, foreign currency and bank deposits were in demand to be held over periods of only a few days. Most credit transactions were made for very short periods, of a month or less. Despite the uncertainty over future prices, these contracts usually specified a nominal interest rate in pesos. For longer-term transactions (such as housing rentals), a common practice was to adjust payments on the basis of monthly price indices.

By mid-1985, public sector prices and wages (both in the government and private firms) were adjusted every month. In general, wage increases seem to have used the CPI inflation rate of the previous month as a reference. Prices set by private firms often changed at intervals of only few days. In hyperinflations, it seems that price setting gravitates to a "dollar standard" or similar systems (see Costantino Bresciani-Turroni, 1937, p. 136). This did not happen in Argentina; although price shifts were increasingly synchronized, there was no single price that governed the movements of the whole set.

Inflation had strong real effects. Although the poor performance of the economy (or the recession that preceded the program) cannot be attributed exclusively to the very high inflation, it is clear that it seriously disturbed economic activities. In any case, by June 1985, both the public and the government saw inflation as perhaps the most urgent of Argentina's economic problems.

II. The Program

Most hyperinflations have ended abruptly without the use of price controls. According to equilibrium theories, a credible fiscal announcement sufficed to stabilize expectations, causing prices to stop immediately (Thomas Sargent, 1982). Other theories, however, doubt that this mechanism operated, and stress rather the stabilization of a "key" price like the exchange rate (Francisco Lopes, 1984; Rudiger Dornbusch, 1985). In the Argentine case, "dollarization" was not complete. It seemed reasonable to assume that there remained significant inflationary inertia (Roberto Frenkel, 1984), both because of the widespread use of indexing with lags and also because of the difficulty in automatically coordinating the stabilization of individual prices. There was a clear risk that fiscal and monetary measures by themselves would not be credible (since real tax revenues depend on inflation), and would have induced a large drop in output as well as erratic price movements. On the other hand, previous experience had shown that income policies have little effect, or soon lose it, if the growth in nominal demand is not clearly under control.

The Austral Plan aimed at a sudden disinflation. A shock approach was chosen because of the urgency of the situation and because a long, persistent disinflation with tolerable output costs would be very difficult to manage. The program had three main parts. First, the government announced that the central bank would stop granting credits to the Treasury. The fiscal deficit would fall drastically. Additional revenues were expected from several sources: the lower inflation itself, higher duties on foreign trade, a forced loan based on direct taxes and the previously decided increases in public sector prices. Government real wages and investment had been falling, and were not expected to rise in the near future. Furthermore, the drop in nominal interest rates would reduce payments on bank reserves (the "quasi-fiscal" deficit of the central bank). Second, prices and wages were frozen. Third, the austral was introduced as the new unit of account. Currency and demand deposits were converted on June 14 at a ratio of 1000 pesos to 1 austral. Subsequent payments denominated in pesos (or indexed with prereform prices), resulting from previous contracts, would be made in australes at a conversion rate that changed daily, in such a way as to make the peso depreciate relative to the new currency. This conversion tried to avoid wealth transfers due to the abrupt disinflation, by compensating approximately the pre- and postreform inflation differential (see Axel Leijonhufuud, 1984).

The program was designed to be robust, in the sense that is did not depend on a single mechanism to act upon prices. The three main parts of the plan were meant to support one another and to provide signals to various segments of the public. At the same time, the program "overdetermined" the economy: the passage from the transition phase (which was viewed as a first priority) to a more permanent regime would entail the removal of some constraints, with difficult decisions to be taken as to the choice of instruments and their management.

III. The Effects

Between June 1985 and March 1986 the CPI increased at an average rate of 3 percent per month. This inflation, although clearly different from zero, was very low by previous standards. The price controls, moreover, did not produce noticeable disturbances in supply. Much of the increase in consumer prices was accounted for by primary goods and services; industrial prices, that had risen very fast before the start of the program, stayed almost constant. Given the lag between the accrual and the spending periods of earnings, it seems that unit real wages in the private sector initially increased, then fell to about their original values. By the end of 1985, many firms apparently started to adjust wages more or less in proportion to the observed growth in the CPI.

There has been much debate about the size of the fiscal deficit, mainly referring to the treatment of central bank accounts. However, the government's borrowing requirements dropped sharply.[2] The central bank effectively stopped extending credits to the Treasury, except for the transfer of foreign loans. Nevertheless, the money supply expanded rapidly, due to the increase in foreign reserves (originating both in a large trade surplus and in capital inflows) and in rediscounts, partly compensated by higher reserve requirements. At the same time, there was a large shift in real money demand. Nominal interest rates fell noticeably, but remained very much above the observed inflation rate in the initial months.[3]

Real output continued to fall immediately after the program started. However, demand and production soon began to recover rapidly. In the fourth-quarter 1985, manufacturing GDP was already higher than in the first quarter, when the recession had not yet fully developed. The reduced price instability probably caused changes in behavior; for example, it is likely that the revival of consumer credit was one of the reasons for the recovery in demand.

The incipient stabilization, however, was not consolidated. The residual inflation apparently put a floor to expectations and amplified the pressures for wage increases. There was much public concern about the possibility of a jump in prices at the end of the freeze. The fiscal situation, although much improved, still looked fragile: some of the initial measures were transitory (like the higher export taxes, when international prices were falling) and there were strong demands for larger expenditures.

Thus, the government found it difficult to aim for a very low inflation. In April 1986, the authorities announced more flexible income policies. The exchange rate and public sector prices increased by small percentages, and it was announced that they would be adjusted periodically. The price freeze was replaced by a system of administered prices, concentrating on large firms. Wages would increase by 8.5 percent in the second quarter; trade unions and firms were called to discuss a new set of basic wages which would validate the previous drift.

Contrary to some anticipations, there was no price shock after the freeze. But the CPI inflation accelerated to 4.5 percent per month in the second quarter. The upward drift became general, so that relative prices varied less than in the previous period. Industrial real wages increased: negotiations between unions and firms apparently brought about

[2] Official estimates of the PSBR (excluding the central bank but including the revenue from the forced loan) indicate a fall from 8 percent of the GDP in the first half of 1985 to around 3 percent between June 1985 and June 1986.

[3] It is difficult to distinguish between the effects of persistent inflationary expectations and insufficient monetization. Dornbusch (1986) argues that the second factor was particularly important.

significant effective raises. The monetary expansion now induced a fall in interest rates, not only in real but also in nominal terms. The growth in aggregate demand allowed the rapid industrial recovery to continue, although firms seemed reluctant to engage new workers or expand capacity.

In July, the CPI increased by 6.8 percent. This created great uncertainty. The parallel ("black market") exchange rate suddenly took off, after having remained almost constant in the previous months. Also, many firms raised their prices. The government reacted by tightening monetary and income policies. Interest rates rebounded and, apparently, industrial sales started to drop; by contrast, the CPI growth rate fell to around 5 percent in November 1986, down from almost 9 percent three months before.

IV. Some Concluding Remarks

The results of the program show that fiscal policies combined with measures to attack inflationary inertia can deal with a very high inflation without causing large real costs. Difficult questions remain regarding the set of policies that may guide prices along a relatively stable path in a country with a long inflationary history, and where distributive conflicts often lead to strong pressures on the budget and to inconsistent price-wage decisions. The heavy foreign debt and declining terms of trade do not make it easier to reconcile the claims of the various groups. There is much latent instability, which demands day-to-day policy management. Still, high inflation and stagnation are more than occasional problems in Argentina. Longer-run changes in the working and financing of the public sector and in the system of economic incentives seem necessary to overcome them. The stabilization program has stimulated a debate within the country that may, hopefully, produce some agreement on such reforms.

REFERENCES

Bresciani-Turroni, Costantino, *The Economics of Inflation*, London: Allen and Unwin, 1937.

Dornbusch, Rudiger, "Stopping Hyperinflation: Lessons from the German Experience in the 1920s," mimeo., MIT, 1985.

______, "Tight Fiscal Policy and Easy Money: The Key to Stabilization," mimeo., MIT, 1986.

Frenkel, Roberto, Salarios industriales e inflación. El período 1976–1982," *Desarrollo Económico*, October-December 1984, *95*, 387–414.

Heymann, Daniel, *Tres ensayos sobre inflación y políticas de estabilización*, Buenos Aires: CEPAL, 1986.

Leijonhufuud, Axel, "Inflation and Economic Performance," in B. Siegel, ed., *Money in Crisis*, Cambridge: Ballinger, 1984.

Lopes, Francisco, "Inflaçao inercial, hiperinflaçao e desinflaçao: notas e conjeturas," mimeo., Pontificia Universidade Catolica do Rio de Janeiro, 1984.

Sargent, Thomas, "The End of Four Big Inflations," in R. Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1982.

Brazil's Tropical Plan

By ELIANA A. CARDOSO AND RUDIGER DORNBUSCH*

On February 28, 1986, with inflation at 400 percent per year, Brazil embarked on her second major stabilization effort in twenty-five years. This paper highlights the institutional features of inflation and contrasts the two stabilization efforts.

The interaction of supply shocks and indexation are the main elements in generating acceleration of inflation in Brazil. Although, large budget deficits in 1959–64 and in 1979–85 supported the inflation process, assigning them more than an accommodating role would mean neglecting the contribution of the supply side to inflation. The Brazilian inflation process is highly institutional. It does not resemble hyperinflations where pricing and wage setting are geared to the exchange rate by the hour, making it possible to stop inflation by containing money creation and fixing the exchange rate. The design of an appropriate stabilization mix, therefore, needs to recognize backward-looking indexation of wages, bonds, and rents. The two stabilization episodes demonstrate that an incomes policy is an essential ingredient to nonrecessionary stabilization. They also show that demand restraint is inevitable if disinflation is to be viable. The 1964 program was gradualist and relied on the supply side on wage repression. The 1986 plan was a heterodox shock treatment centered around an uncompromising price freeze and paying insufficient attention to the need for fiscal restraint.

I. A Puzzle

In high inflation economies, institutional arrangements provide for a periodical resetting of wages. The peak real wage occurs at the time of the nominal wage increase. Subsequently, up to the next adjustment, the real wage is eroded by inflation. Figure 1 shows the real value of the minimum wage in Brazil over the past ten years.

Escalation of inflation invariably involves a shortening of adjustment intervals for wages, which in turn further increases inflation: in a context of overlapping contracts, the shortening of the intervals increases the number of wages revised on the same date thus pushing up costs. In 1979, the annual adjustments of wages in Brazil shrunk to a twice-yearly base and the inflation rate doubled. By the end of 1985, wages were beginning to move into 3-month revision cycles. The government, keenly aware that the transition to even shorter periods must have hyperinflationary consequences tried to stop this process.

We distinguish the payments period, which is weekly or monthly, and the frequency of inflation adjustment. Wage adjustment for past inflation occurs at fixed intervals. In any inflation process, intervals shrink to one-year, 6-month, 3-month, 1-month, and ultimately to the daily course of the dollar. The puzzle is why adjustment intervals show so much inertia. The average real wage during the interval depends on the rate of inflation and the length of the interval. Given the expected inflation rate, the same average real wage can be obtained by the combination of shorter intervals and lower real peaks, or larger intervals and bigger peaks. If wage earners are unable to lend and borrow in perfect capital markets, they are forced to hold cash in order to sustain their consumption levels during the later part of the cycle when their real wages are below the average. Cash held for later consumption has its purchasing power eroded by ongoing inflation.

Wage earners who do not have access to perfect credit markets clearly should prefer

*Associate Professor, Fletcher School of Law and Diplomacy, Tufts University, and Professor of Economics, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 01239.

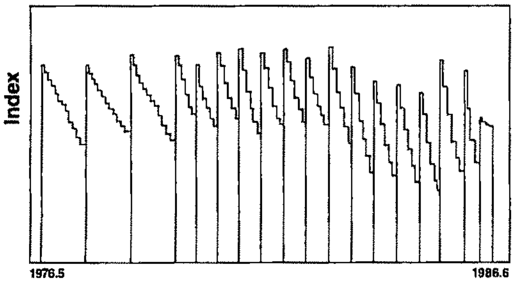


FIGURE 1. MINIMUM REAL WAGE MONTHLY DATA

Source: Conjuntura Economics.

an even flow or real wages rather than the sawtooth pattern of Figure 1. They must be willing to trade smaller nominal wage increases for shorter periods of wage resetting. The costs of more frequent inflation adjustment of wages for firms are negligible. Adjustments of the pay period use resources, but not the indexation of wages by some commonly public inflation index. Yet we only observe a shrinking of intervals when the inflationary erosion becomes extreme.

II. Other Institutional Aspects of Inflation

Stanley Fischer (1977) and John Taylor (1979) have drawn attention to the persistence of price disturbances in a setting of overlapping wage contracts even under rational expectations and a well-understood program of monetary control. In the Brazilian setting, institutional factors take, to a large extent, the place of relative wage and expectations mechanisms that characterize Fischer-Taylor contracts. Mandatory indexation is backward-looking and periodically readjusts wages and other prices. We express this inflation process in equation (1). Wage inflation, w_t, is equal to past price inflation plus a disturbance term. *Ceteris paribus*, current inflation, p_t, is equal to past inflation, via indexation of wages as well as public sector prices and the exchange rate. The output gap affects current inflation because it influences the marginal labor cost of firms since turnover of the labor force can be used for wage cutting. The third term, u_t, represents supply shocks. These are superimposed on indexation which propagates them.

$$(1)\quad p_t = w_t + a\, gap_t + u_t; \quad w_t = p_{t-1} + v_t.$$

This inflation process has several implications. First, current supply shocks are automatically transmitted to future periods. Oil price increases, real depreciation, indirect taxes, elimination of public sector subsidies, or agricultural price shocks raise the current rate of inflation and, via indexation, inflation in subsequent periods. In fact, to raise public sector real prices or cut real wages in the presence of full indexation, the frequency of adjustment of exchange rates and public sector prices has to be higher than the frequency of wage adjustments. Only then is it possible to beat the indexation. Indexation of the financial system, of the tax structure, and of the public debt imply that changes in the inflation rate are automatically and fully accommodated.

Second, a slowdown in the growth rate of nominal spending cannot eliminate inflation from one day to the next. Demand restraint runs counter to the automatic cost-inflation that comes from lagged inflation captured by the wage inflation term. The neoclassical answer of instant recontracting of the labor force with reduced wage adjustments in the face of a shift to a noninflationary monetary regime is implausible. The presence of *inertia* is thus one good reason for the use of incomes policy in a stabilization program.

Third, those contracts which are not explicitly indexed in a backward-looking way as, for example, short-term loans in the financial system, carry forward-looking inflation adjustments. Their maturity may run as far as a year. A sudden disinflation would imply an arbitrary redistribution between debtors and creditors.

Fourth, any escalation of inflation started off by some supply shock encounters further endogenous elements that feed the inflation process. One is the increase in the velocity of money. Another one is the inflationary erosion of tax revenue which then implies increased rates of monetary expansion. And still another one is the endogenous shortening of the interval for inflation adjustment of wages, public sector prices, and the exchange rate.

III. Two Stabilizations

The main reason for the sharp increase in inflation between 1959 and 1962 was the fast increase in demand. Between 1957 and 1962, industrial output grew at 11 percent per year. The share of the central government budget deficit in output increased from 2.8 percent in 1960 to 4.3 percent in 1963, while the seignorage share in gross domestic product widened from 3.6 percent in 1959 to 5.7 percent in 1962. The combination of a 30 percent deterioration of the terms of trade, the lack of external finance, a bad coffee crop in 1963, and an agriculture disaster in 1964 all contributed to the inflation problem. The economic crisis was the vehicle for a military take-over on March 1964. The *Programa de Ação Econômica do Governo* (*PAEG*, 1964/66) detailed a plan to reduce inflation gradually in three years using fiscal consolidation and incomes policy. Fiscal consolidation led to a gradual reduction in the deficit from 4.2 percent of gross domestic product in 1963 to only 1 percent in 1966. The main instruments of this budget balancing were increases in public sector prices, cuts in subsidies, and increased tax collections.[1]

Disinflation was achieved by breaking the link between current and past inflation: wage adjustments were made forward-looking and limited to an officially imposed inflation forecast. The cut in wage inflation helped absorb the impact of public sector price increases and exchange depreciation. But the reduction in price inflation fell by a wide margin short of the decline built into wage agreements. Real wage cuts made room both for budget balancing and for an improved external competitiveness.

Fiscal and credit incentives were given to firms which would accept a convenant not to raise prices by more than a stated percentage. Indexation in financial markets was used to mobilize domestic saving and to create a market for public sector debt. Th black market premium that had reached 60 percent in the last quarter of 1963, by the end of 1964 was down to near zero and remained there in 1965 and 1966.

TABLE 1—BRAZILIAN MACROECONOMIC DATA: 1982–86

	1982	1983	1984	1985	1986[a]
Inflation[b]	98	142	197	227	60
Growth	0.9	−3.2	4.5	8.3	7.0
Deficits[c]					
Budget	16.7	19.9	22.2	27.1	9.9
Operational	6.5	3.0	2.7	4.3	4.1
Current Account	8.5	3.5	0	0.1	0.5
Interest Payments	6.5	5.3	5.4	4.7	3.7

Source: Banco Central do Brasil.
[a] Estimate.
[b] December–December.
[c] Percent of gross domestic product.

Inflation declined from 144 percent in the first quarter of 1964 to 57 percent in 1965 and 38 percent in 1966. Industrial production declined in the first year of stabilization by 5 percent, but by 1966 it was already 6 percent above the pre-crisis level. In 1968, a new plan was adopted: it introduced a crawling peg exchange rate, made credit more abundant, and revised the wage adjustment rule. The 1965–68 reforms were the basis for an extended period of strong growth with stable inflation.

The next inflation escalation started with the second oil shock and the shortening of the intervals for wage setting in 1979. It accelerated with the large real depreciation in 1983, with an agricultural disaster and with correction of prices of the public sector and subsidies cuts (see Table 1). In 1985, the new democratic government embarked on a program of expansion which, combined with another bad crop, led to a shortening of the interval for inflation adjustment in some sectors of the private sector to only 3 months.

Mindful of inflation acceleration and upcoming elections, the government embarked on a program of stabilization. The key point of the *Cruzado Plan* was to eliminate catch-

[1] Statistical data is from *Conjuntura Economica*, various issues, *Brasil*: *Programa Economico*, various issues, and Raymond Goldsmith (1986). For references on stabilization programs in Brazil, see Dornbusch and Mario Simonsen (1987).

up inflation in wages and rents. Wage contracts with future readjustment were rolled up and contracts that had experienced a recent readjustment were rolled back. The minimum wage received a bonus of 15 percent increase over its past real average and other workers an 8 percent bonus. A 20 percent cumulative inflation threshold (*escala movel*) was introduced for wage adjustments. All prices and the exchange rate were frozen until further notice. A *tablita* was devised to eliminate the expected inflation built into extant contracts and thus avoid arbitrary redistribution between debtors and creditors. A new currency was introduced to help facilitate the readjustment. On the fiscal side, the tax reform of December 1985 was expected to close a sizeable budget deficit. But tax revenues rose disappointingly due to a lowered income tax withholding schedule and an increased reliance on taxation of financial assets which were no longer popular.

External factors favored the program. The decline in world interest rates reduced the debt-service burden. Sharply lower world oil prices reduced the import bill while the dollar depreciation helped achieve a gain in competitiveness.

There was a sharp initial monetization of the economy. In the first 3 months following stabilization, the monetary base doubled. Between February and July, industrial production increased by more than 10 percent relative to the same period a year before while cumulative inflation was close to zero. Fuelled by strong popular support, policymakers elevated the price freeze to a national fetish. But the budget was allowed to deteriorate. Revenues of state-owned companies were hurt by the price freeze, subsidies cut during 1983–84 staged back and the public sector wage bill increased in line with the economywide trend. The trade surplus disappeared, and shortages and black markets became pervasive. Adjustments made in August went in the direction of very special excise taxes: so large that they were claimed to solve the budget problem, and so small that they could be eliminated from the official price index. A second round of such excise tax increases was imposed in November. Once again their effect was eliminated from the price index. Clearly the government was trying to reenact the 1964 program of real wage cuts to restore the external balance and the budget. But a democratic regime puts severe limits on such exercise.

There is yet another important difference between the two programs. By 1964, the preceding high inflation with no indexation had reduced the real value of the public debt to less than 4 percent of gross domestic product. In the 1980's, by contrast, prevailing indexation and high real interest rates had left a debt-income ratio (foreign and domestic debt combined) of 50 percent of gross domestic product as a mortgage for stabilization. Insufficient budget improvement and the wage-induced boom pushed up nominal interest rates and the black market premium. By the end of 1986, the black market for dollars stood at a premium of more than 100 percent and the short-term interest rate reached 150 percent. Inflationary expectations were becoming extreme. The removal of indexation meant that inflation could accelerate dramatically.

IV. Issues and Lessons

The two Brazilian stabilization programs teach a number of lessons. First, incomes policy is a valuable means in achieving disinflation. It helps avoid dramatic unemployment. But incomes policy by itself is not enough.

Second, disinflation is not costless. In 1964, wage repression was the front payment for disinflation. In 1986, it was the redistribution from firms to workers implicit in the rise in real wages, a loss in exchange reserves, and a deterioration in competitiveness.

The third lesson concerns indexation. Indexation in the presence of supply shocks propagates inflation. But it also protects an inflation rate against rapid acceleration. Between 1965 and 1968, indexation was reinforced and broadened. In 1986, by contrast, it was eliminated altogether and replaced by adjustment triggers without cap. This setting has led to a highly volatile atmosphere where the absence of financial indexation drives asset holders to goods and to the black market and where inflationary expectations can easily become the driving mechanism for an actual inflation process. Inflation now, contrary to the past twenty years, can become self-generating. The lesson is that restoration of indexation in labor and asset markets, with long adjustment intervals, is good advice.

Fourth, following disinflation, real interest rates turn sharply positive unless the government engages in a significant monetization. It is difficult to know what is enough because high inflation partially destroys traditional linkages between interest rates and real balances. But being too conservative is problematic, because high real interest rates in the presence of a large public debt create a fiscal problem.

A final point concerns the budget. The possibility of financing a small deficit in a noninflationary manner depends on the growth rate of output, the prospective path of real tax revenues, and the real rate of interest. If output growth is high and the real rate of interest is negligible, there is room for deficits. If the relation is the reverse, there is the risk of building up a fiscal problem which ultimately requires inflationary liquidation.

REFERENCES

Dornbusch, Rudiger and Simonsen, Mario Henrique, "Inflation Stabilization with Incomes Policy Support," *Group of Thirty*, New York, 1987.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Goldsmith, Raymond, *Brasil, 1850–1984: Desenvolvimento Financeiro sob um Século de Inflação*, Brasil: Harper & Row, 1986.

Taylor, John, "Staggered Wage Setting in a Macro Model," *American Economic Review Proceedings*, May 1979, *69*, 108–13.

Conjuntura Economica, Fundação Getúlio Vargas, various issues.

Brasil: Programa Economico, Banco Central, various issues.

Economic Conditions and Gubernatorial Elections

By SAM PELTZMAN*

The literature on the effect of economic conditions on election outcomes has so far focused mainly on presidential (Ray Fair, 1978; Allan Meltzer and Mark Vellrath, 1975; Burton Abrams and Russell Settle, 1978) or congressional elections (Gerald Kramer, 1971; George Stigler, 1973; Howard Bloom and H. Douglas Price, 1975). Typically some aggregate voting measure (for example, the incumbent party's share of the national vote) is regressed on macroeconomic measures of the putative success or failure of the incumbents' economic policy. The broad consensus of the literature is that voters reward "good" economic performance, but that they have short memories. Usually income growth in the year preceding the election dominates measures like the unemployment rate or inflation rate, but prior economic performance seems not to matter at all.[1]

The existing literature is constrained to work with small samples. For example, there have been fewer than 20 presidential elections since countercyclical monetary/fiscal policy became important. The small sample compromises the power of any test, or, where pre-1930's data are used, raises questions about the sophistication of either the voters or the research design. This paper is a modest attempt to overcome the small sample problem by focusing on post-World War II gubernatorial elections, of which there have been several hundred.

These extra degrees of freedom are not really free, because governors can have only limited effects on the economic welfare of voters. In the organization chart of the American federal system, governors and presidents share similar powers of appointment, budget making, etc., and the role of the governor's mansion as a training ground for presidential candidates is well established. But, as chief executive in a small open economy without a central bank, the governor cannot conduct very powerful macro policy. That weakness does yield a benefit: We can begin to see if voters make sensible connections between policy and performance. That issue was raised by Stigler, who argued that sensible voters would ignore short-run economic fluctuations, but it has since been largely ignored. How then are we to interpret the result that voters' reward good short-term macro performance when this result comes from time-series including pre-New Deal or even pre-Federal Reserve data (as in Fair or Kramer)? If early presidents, like today's governors, could not plausibly "perform," the inference would appear to be that voters are not very sensible.[2] The subsequent evidence shall, however, cast some doubt on that inference.

I. Data and Model

I use a sample of 269 postwar elections which comprises essentially all gubernatorial

[†] *Discussants*: Assar Lindbeck, Institute for International Economic Studies, Stockholm; James Snyder, University of Chicago.

*Sears, Roebuck Professor of Economics and Financial Services, Graduate School of Business, University of Chicago, Chicago, IL 60637. I thank David Barker for research assistance, and the Bradley Foundation and the Center for the Study of the Economy and State, University of Chicago, for financial support.

[1] This result has motivated research on political business cycles, in which politicians respond to the voters' short memories with stimulative policies just prior to elections.

[2] I eschew the term "rational," since voters may be rationally nonsensical, given their putative lack of incentive to invest in information on the connection between policy and performance.

elections to four-year terms from 1949 through 1984 in states with competitive party systems.[3] I then follow the literature and try to connect the incumbent party's share (IPS) of the two-party vote to economic performance (EP) during the governor's term. I assume that the representative voter has some normal probability (K_i) of voting for the incumbent party from which he deviates according to his estimate of the impact of the party's policies on his economic welfare (EW_i). Aggregating over voters, we get

$$(1) \qquad IPS_t = K + f(EW_t),$$

I allow voters to translate EP into EW slowly according to

$$(2) \quad EW_t = (1-w)EP_t + w \cdot EW_{t-1}, \qquad 0 < w < 1.$$

That is, voters remember their estimate of EW at the last election and give it some weight (w) in calculating EW_t. A linear version of (1) which incorporates (2) is

$$(3) \quad IPS_t = K(1-w) + m(1-w)EP_t + w \cdot IPS_{t-1},$$

where m and w are (assumed) constant. To implement (3) on the sample, I allow K to vary across states, but not over time. So, operationally, K is the incumbent party's long-run average share in a state. I also allow for the well-known advantage to incumbents seeking reelection. An implication of this scheme is that the performance measures used by sophisticated voters should be "unexpected" at $t-1$; any component of performance which is predictable at $t-1$ should be reflected in that election's IPS.

A pervasive result of my preliminary work was that voters do respond consistently to the surprises in performance rather than to the expected component. For example, cursory examination of annual real income, inflation, and unemployment series reveal considerable persistence, while growth of real income, change in inflation, and change in unemployment are approximately random walks, and these latter are what voters react to. Another pervasive result is reaffirmation of the voters' short memory. This shows up in two ways: 1) no estimate of w was ever far from zero (see Table 1), and no performance measure going back more than a year or so before election day (with one notable exception) ever "mattered." Accordingly subsequent results focus on surprises in the election year.

II. Results

Regression (1), Table 1, is the obvious extension of the literature to these data. The incumbent party's vote share is regressed, inter alia, on the growth of real per capita income[4] in the state in the election year. This regression hints that good performance *hurts* the incumbent party, but any paradox vanishes on closer inspection. State income growth, of course, reflects fluctuations in national income, and it turns out that these are the dominant influence on gubernatorial elections. The (marginal) voter apparently understands that governors have little influence on the growth of state income, and he or she accordingly rewards (penalizes) the *party* of the incumbent *president* for good (bad) macro performance. This is shown in regression (2). Here performance is measured by national per capita income growth, and slope and intercept dummies are included which distinguish whether the incumbent party in a state is the same (= +1)

[3] A shrinking number of states have two-year terms, and I excluded these elections. When a state adopts a four-year term, I included that state from the second four-year term election. I also exclude Alaska and Hawaii and all elections in states where a single party received over 70 percent of the vote for at least three consecutive elections. In effect this criterion eliminates a few southern states where Democrats have received only token opposition for some or all of the period. Over 90 percent of all gubernatorial elections occur in even numbered years.

[4] Nominal per capita personal income in the state deflated by the national personal income deflator.

TABLE 1—REGRESSIONS OF INCUMBENT PARTY'S SHARE OF VOTE;
269 GUBERNATORIAL ELECTIONS, 1949–84

Independent Variables	Coefficients/ *t*-Ratios		Mean
(in % except for dummies)	(1)	(2)	S.D.
Incumbent Party Share	−.01	.11	55.7
in Last Election	(0.1)	(1.4)	(5.4)
Incumbent Candidate	4.4	4.0	.56
= +1, 0 otherwise	(4.7)	(4.6)	(.50)
Per Capita Income Growth in			
(a) State	−.16		1.7
	(1.2)		(3.4)
(b) Nation		−.33	1.9
		(1.8)	(2.3)
(c) State-Nation		.03	−.25
Presidential Dummy = +1 if		−4.1	.25
Incumbent Party is same as		(5.6)	(.97)
President, −1 if different			
Presidential Dummy×			
National Income Growth		.86	.70
		(3.1)	3.0
			Incumbent Party Share
$\bar{R}^2$	.24	.34	52.6
SEE	7.0	6.5	8.0
	(3)	(4)	
Incumbent Party Share	.12	.08	
in Last Election	(1.5)	(1.0)	
Incumbent Candidate	4.1	4.1	
	(4.7)	(4.7)	
National per Capita			
Income Growth	−.32	−.22	
	(1.7)	(1.2)	
Change in Inflation	.08	.15	.09
Rate	(0.4)	(0.8)	(2.2)
Growth of State Revenue/		−.08	10.0
State Income		(2.6)	(13.9)
Presidential Dummy	−4.0	−4.0	
	(5.4)	(5.5)	
Presidential Dummy×			
(a) National Income Growth	.82	.86	
	(2.9)	(3.1)	
(b) Change in Inflation	−.29	−.29	.00
	(1.5)	(1.6)	(2.2)
$\bar{R}^2$	.34	.36	
SEE	6.5	6.4	

Sources: Political data: Richard Scammon and Alice McGillivray, *America Votes* (various issues); Economic data: U.S. Bureau of Census, *Statistical Abstract of the Unites States* and Council of Economic Advisers *Economic Report of the President*, various issues. *Notes:* See text for definitions of variables. *T*-ratios are shown in parentheses below coefficients. Each regression includes a vector of state-party dummy variables, but neither their coefficients nor the constant term is shown. This vector represents *K* in (3). The dummy for each state (or combination of neighboring states if a state has fewer than five elections in the sample) = +1 when the incumbent party in the state is Republican, −1 when Democrat, so the dummy vector allows for long-run average differences in party strength among states.

or different (= −1) from the president's party. The significantly positive coefficient of the slope dummy (+.86) is a measure of the reward for good macro performance bestowed on the gubernatorial candidate from the president's party.[5] The regression also suggests a curious asymmetry: the negative coefficient of income growth (−.33) implies that the reward is greater if the president's party is challenging rather than defending the governor's mansion.[6] Regression (2) also implies that, once national income growth is accounted for, the difference between state and national growth doesn't matter. This result survives all subsequent refinements.[7]

Regression (3), Table 1, adds the change in the inflation rate (I use the Personal Income deflator) and its interaction with the presidential party dummy. The previous research has used the inflation rate and obtained mixed results.[8] I found it insignificant in a counterpart to regression (3) not shown. However, there is considerable persistence in annual inflation rates,[9] which essentially disappears on first differencing. Thus, if voters dislike inflation, it makes no sense for them to, for example, penalize an absolutely high inflation rate if it is declining. Indeed, the change in inflation seems to be the more relevant inflation variable. The negative coefficient of the presidential party change in inflation interaction suggests that the president's party is penalized for unpleasant surprises in inflation.[10]

[5] I was also able to confirm that income dominates unemployment as a performance measure. Substituting the change in the unemployment rate for income growth produced qualitatively similar but perceptibly weaker results than regression (2).

[6] Specifically, when the president's party is defending the gubernatorial mansion, a 1 percent acceleration of income growth increases his party's vote share by .53(= .86 − .33) percent. When the president's party is challenging, the increase is +1.19(= .86 + .33) percent.

[7] That is, the difference between state and national growth is never significant when added to any subsequent regression. Neither is the interaction of this difference and the presidential dummy. So voters do not penalize the candidate of the president's party if local income grows more slowly than national income.

[8] Fair found the inflation rate insignificant. Kramer found it significant after correcting a data error in his original article (Saul Goodman and Kramer, 1975).

[9] The first-order serial correlation of the inflation rate is +.81 over my sample period.

[10] The insignificant positive coefficient of the change in inflation hints weakly at the same sort of asymmetry found for income growth.

Governors do have one policy lever that can perceptibly affect voters' economic welfare, namely their influence on the state budget. Regression (4), Table 1, adds the *four-year* percentage change in state general revenues divided by state personal income, and it suggests that voters penalize budgetary expansion. In this case voters' memories seem longer than for national performance; the one-year value for this variable proved insignificant. Because state revenues come from federal aid as well as taxes, I tried a refinement which distinguished the local from the federal aid component of revenues. But this decomposition did not improve on regression (4);[11] voters do not act as if federal aid is a free lunch. Since both revenues and federal aid have, until recently, been rising relative to income, these results raise obviously intriguing questions of interpretation.

III. Summary

Voters in gubernatorial elections seem able to draw appropriately delicate distinctions in their response to economic conditions. They vote as if they understand that national rather than local policies have the dominant effect on their income. They also act as if they understand that national policies affect national income more than its geographic distribution. They do this, on the evidence here, by holding gubernatorial candidates of the president's party hostage to the perceived effectiveness of his macro policy and by ignoring local idiosyncracies. Voters also will get good grades from economists for correctly distinguishing expected from unexpected inflation. Finally, voters do respond to the local variable that the governor can control: they penalize growth of the state budget.

My results may offer a clue about voters' seeming myopia. About two-thirds of gubernatorial elections coincide with the midterm congressional election. If (the voters believe) the president is the responsible agent for macro policy, these midterm elections give voters an opportunity to "settle up" with the president for the past two years. The signal that emerges should then alter or affirm macro policy, and the president's response to it would be monitored at the next presidential election. On this view, voters should respond to performance in each biennium, and, given the usual policy lags, election year surprises would not be a bad proxy for biennial performance.[12]

Direct comparison of my results with the previous literature is complicated by the asymmetry I find between challengers and defenders. To facilitate comparisons, consider a stylized case in which the president's party is defending half the contested governor's mansions and challenging the other half. Then each extra percentage point of national income growth increases the average vote share of the candidates from the president's party by over $\frac{3}{4}$ of a percentage point (regression (4)). This is between Kramer's result for congressional candidates (about $\frac{1}{2}$ percent) and Fair's for presidential elections (about 1 percent). A similar calculation for inflation yields an extra $\frac{1}{3}$ point in vote share for each point of deceleration in inflation, about the same as Fair's (insignificant) result for the inflation rate itself. Some idea of the importance of these effects may be gleaned from the data in the last column of Table 1. The standard deviations of the growth and inflation variables are both around $2\frac{1}{4}$ percent. So, a simultaneous swing from 1 standard deviation below to 1 standard deviation above the mean of these variables would add about 5 points to the average vote share of party confreres of the president. Many election outcomes will sur-

[11] Specifically, since revenue/income = state's own revenues per dollar of income × (1 + federal aid/own revenues), I entered the growth rates of both components separately. Each had indistinguishably different negative coefficients. Substituting expenditures for revenues produces results similar to those in (4).

[12] In fact, preliminary results imply that there is little to choose statistically between one- and two-year performance measures.

vive such a swing, given the large standard deviation of the incumbent party share. However, about half these shares lie between 45 and 55 percent, where performance swings could conceivably make a difference. By comparison, it takes very large tax increases —about at the maximum of the revenue growth variable in my sample—to cost the incumbent party 5 vote share points.

Finally, it is worth asking what is gained from the extra detail provided by local election returns. Those of my results which can be compared with past analyses of aggregate time-series are not spectacularly different or stronger. So the answer has to focus on refinements. For example, my results deepen the puzzle about voters' myopia by showing that it is national rather than local income that counts; voters are not merely turning the rascals out when their last paycheck is reduced. Neither this nor the added puzzle of the asymmetry between challengers and defenders could plausibly have been revealed by the less detailed data. If we lack a theory capable of rationalizing such puzzles, the larger sample at least provides the flexibility to delineate them more precisely.

REFERENCES

Abrams, Burton and Settle, Russell, "The Economic Theory of Regulation and Public Financing of Presidential Elections," *Journal of Political Economy*, April 1978, *86*, 245–57.

Bloom, Howard and Price, H. Douglas, "Voter Response to Short Run Economic Conditions: The Asymmetric Effect of Prosperity and Recession," *American Political Science Review*, December 1975, *69*, 1240–54.

Fair, Ray, "The Effect of Economic Events on Votes for President," *Review of Economics and Statistics*, May 1978, *60*, 159–73.

Goodman, Saul and Kramer, Gerald, "Comment on Arcelus and Meltzer," *American Political Science Review*, December 1975, *69*, 1255–65.

Kramer, Gerald H., "Short Term Fluctuations in U.S. Voting Behavior, 1896–1914," *American Political Science Review*, March 1971, *65*, 131–43.

Meltzer, Allan and Vellrath, Mark, "The Effects of Economic Policies on the Vote for the President," *Journal of Law and Economics*, December 1975, *18*, 781–98.

Scammon, Richard and McGillivray, Alice, *America Votes*, Washington: Elections Research Center Congressional Quarterly, various issues.

Stigler, George, "General Economic Conditions and National Elections," *American Economic Review Proceedings*, May 1973, *63*, 160–67.

U.S. Bureau of the Census, *Statistical Abstract of the United States*, Washington, various issues.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, various issues.

The Revealed Preferences of Political Action Committees

By KEITH T. POOLE, THOMAS ROMER, AND HOWARD ROSENTHAL*

Massive campaign spending in recent congressional elections has sparked concern with the activities of Political Action Committees (PACs), which have accounted for a large and growing share of campaign finance. Recent empirical work has examined the impact of campaign spending on electoral outcomes (Gary Jacobson, 1985), the effect of contributions on legislative behavior (John Wright, 1985, provides an overview of results), and the contribution patterns of individual PACs or groups of PACs (J. David Gopoian, 1984; Poole and Romer, 1985). While money does not guarantee election, it is most frequently the case that winners outspend losers, and that incumbents in congressional races receive far more in campaign contributions than do their challengers. Spending is highest in races that are expected to be close, even when some account is taken of the obvious simultaneity between closeness and spending. As to the effect of contributions on voting in Congress, the results so far have been quite tenuous. There is little systematic evidence on whether votes in Congress are influenced by campaign contributions, though the journalistic presumption of influence is strong. In this paper, we look at the way PACs allocate their money, taking incumbents' voting records as given.

I. Modeling Considerations

We focus on PAC contributions to candidates in races where an incumbent member of the House of Representatives is running for reelection. Most attempts to "explain" PAC giving as a function of many independent variables obtain results with low explanatory power. This is especially true if one examines spending by *individual* PACs. We believe that explanatory power will be inherently low as a result of the complicated resource allocation problem that confronts PACs. Even if one ignores congressional primaries and other contests, PACs can choose among nearly 470 races for the House and Senate in each two-year electoral cycle. It is hardly surprising then that, in this political supermarket, a given PAC typically contributes nothing to most races. The presence of many zeros in the data renders standard regression analyses of individual PACs inappropriate.

A model of how PACs contribute should take account of the following considerations:

1) A PAC is most likely to contribute to a candidate who is expected to promote policies it favors. Generally, measuring the policy match between a PAC and an incumbent has been ignored or treated in a casual way. For example, the incumbent's rating by the Americans for Democratic Action (ADA) is often used as a proxy for the PAC's evaluation of the incumbent's voting record. Using ADA rating, however, makes sense only if the ADA rating is highly correlated with the PAC's evaluation. In other cases, *ad hoc* vote indices are constructed by the researcher, based on a small number of roll calls that the researcher believes are of interest to the PAC. Fortunately, some organizations that sponsor PACs publish ratings of incumbents. These ratings, expressed on a scale running from 0 to 100, summarize the incumbent's voting record on legislative items that are of greatest concern to the group in each year. These ratings are clearly more appropriate than either type of external index.

It is more difficult to gauge how a PAC evaluates a challenger. In this paper, we assume that a PAC contributes to a challenger when it has a sufficiently negative evaluation of the incumbent. Challenger characteristics do not enter the model.

2) A PAC is more likely to contribute, *ceteris paribus*, in races that are expected to be close. At the margin, a dollar of spending

*Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA 15213.

is more "productive" in a close race than in one that is not highly competitive.

3) A PAC is more likely to give to someone who can be instrumental in favoring its policy goals. This consideration has led researchers to use measures of seniority and committee chairmanship. As Poole and Romer (1985) found that chairmanship was not an important predictor of contributions by aggregate groupings of PACs, we focus solely on seniority.

4) For the PAC's contribution to be influential, in terms of being recognized after the election, it should be substantial. For that reason, we would not expect to find contributions of $5.00. This consideration, coupled with resource limitations, causes most PACs to walk away from most House races. Thus it is important to model how PACs select the relatively few races where they do make a contribution.

II. Data

We consider the 386 House races in 1980 in which an incumbent was running.[1] Of the 31 interest groups that published ratings for the 1979 session of Congress, there were 12 that operated PACs and made contributions in more than 15 of the House races. These 12 PACs are the focus of our analysis. They include seven labor PACs, two general business PACs, and three "ideological" PACs.[2]

Campaign contributions are those reported by the Federal Election Commission for the 1979–80 electoral cycle. For each incumbent, we defined *net money* as the difference between contributions by a given PAC to the incumbent and those to his 1980 challenger. Thus "positive" money indicates support for the incumbent, and "negative" money means support for his challenger. Contributions by a PAC to both sides in a race are extremely rare in our sample, so net money generally indicates clear preference for one side or the other.

III. Who Gets Money

Our approach to modeling the four factors we noted as important in PAC giving is to consider first the choice among the three alternatives of contributing to the incumbent, contributing to the challenger, and not contributing to either. We treat PACs as taking an incumbent's voting record as given, so that PACs are seen as using past voting behavior as a good predictor of future positions. The basic idea is that a PAC evaluates incumbents according to a latent interval scale Y^*, which of course is unobserved by us. We assume that this latent scale can be expressed as a linear function of observable characteristics X plus a random normal error, e:

$$Y^* = X\beta + e. \tag{1}$$

A PAC will make a contribution to incumbent i's campaign if Y_i^* exceeds a threshold value μ_1. If Y_i^* falls below another threshold value μ_0, the PAC will contribute to the incumbent's opponent. A PAC will make no contribution in a race when $\mu_0 < Y_i^* \leq \mu_1$. Let Z_i be an ordinal level variable that is equal to 1 if the PAC made net contributions to incumbent i's campaign; 0 if the PAC made no contributions in the race for i's seat; and -1 if the PAC made net contributions to i's challenger(s). Then

$$Z_i = \begin{cases} -1 & \text{if} \quad Y_i^* \leq \mu_0 \\ 0 & \text{if} \quad \mu_0 < Y_i^* \leq \mu_1 \\ 1 & \text{if} \quad Y_i^* > \mu_1 \end{cases} \tag{2}$$

[1] We excluded the race of O'Neill (D-Mass.) who, as Speaker of the House, does not normally vote, and so is unrated by interest groups. We also excluded two races in which the incumbent was a member who had won a special election, and did not have a complete voting record for 1979.

[2] The labor PACs are those affiliated with American Federation of State, County, and Municipal Employees (AFSCME), American Federation of Teachers (AFT), Building and Construction Trades Department of AFL-CIO (BCTD), Committee on Political Education/AFL-CIO (COPE), National Education Association (NEA), United Auto Workers (UAW), and United Mine Workers (UMW). The business PACs are Chamber of Commerce of the United States (CCUS) and National Federation of Independent Business (NFIB). The others are ADA, Committee for the Survival of a Free Congress (CSFC), and League of Conservation Voters (LCV). Where a group was connected to more than one PAC, we aggregated the contributions made by that group's PACs.

TABLE 1—TRICHOTOMOUS PROBIT RESULTS
($N = 386$ for each PAC)

PAC	Parameter			
	μ_1	β_0	β_1	β_2
1) AFSCME	2.89332	1.05605	0.02356	0.01681
	(17.30)	(6.21)	(7.13)	(3.53)
2) BCTD	2.79627	0.22937	0.04180	0.03964
	(17.30)	(0.80)	(7.28)	(4.19)
3) COPE	2.68510	0.78016	0.01867	0.03039
	(14.32)	(4.67)	(5.13)	(6.61)
4) NEA	3.28964	1.12006	0.02946	0.01186
	(13.02)	(6.70)	(7.71)	(2.74)
5) NFIB	2.69694	−0.21879	0.03700	0.02191
	(12.37)	(−0.97)	(7.67)	(4.20)
6) UAW	2.13674	0.18733	0.03488	0.00593
	(13.22)	(1.26)	(10.67)	(1.32)
7) AFT	3.36606	1.11484	0.01972	0.01587
	(17.36)	(6.33)	(5.48)	(3.19)
8) ADA	5.62655	2.66910	0.01077	0.04197
	(7.87)	(4.82)	(0.75)	(4.21)
9) CCUS	3.37438	1.24182	0.00716	0.02884
	(17.97)	(5.58)	(1.43)	(5.38)
10) CSFC	4.14796	1.08786	0.01934	0.02461
	(10.02)	(3.30)	(2.28)	(3.82)
11) LCV	4.36930	2.24101	0.00615	0.02414
	(19.02)	(4.63)	(0.68)	(2.58)
12) UMW	3.77646	1.66876	0.00932	0.00528
	(14.81)	(7.12)	(1.88)	(0.93)

Notes: The *t*-statistics are shown in parentheses. PACs 1–6 each made contributions in more than 100 races; PACs 7–12 were less active, making contributions in 23 to 81 races.

We estimate the parameters of this trichotomous probit system by maximum likelihood methods (Richard McKelvey and William Zavoina, 1975). In this model, two parameters are not identified. We follow the standard practice of setting σ^2, the variance of the disturbance term e, equal to 1, and the first threshold μ_0 to 0. For each PAC, we then estimate the second threshold μ_1 and the linear parameters β.

In defining X in equation (1), we experimented with a variety of specifications involving ratings, vote margins, and seniority. The estimates reported in Table 1 are based on the following specification of the latent variable Y^*:

$$(3)\quad Y^* = \beta_0 + \beta_1 R + \beta_2 M(R-50) + e,$$

where R is the 1979 rating of the incumbent by the PAC's sponsoring group, and $M=1$ if the incumbent's vote margin over his major-party opponent in 1978 was less than or equal to 25 percentage points; otherwise $M=0$.

The interaction term $M(R-50)$ is intended to capture the notion that giving is more likely in races that are expected to be close.[3] With $\beta_2 > 0$, a highly rated incumbent in a close race has a greater probability of receiving a contribution (i.e., having the latent index be above μ_1) than his equally highly rated colleague who is not expected to face any difficulty. In the case of an incumbent with a low rating (for example, $R \approx 0$), the probability that the PAC will make a contribution to the challenger is greater if the race is expected to be close than otherwise.

Table 1 reports the estimates of the trichotomous probit model when the underlying latent variable is specified as in equation (3). The parameter μ_1 is the value of the latent variable that corresponds to a PAC's decision threshold for making a contribution to the incumbent. Since μ_0 is normalized to 0, the estimated t-statistic for μ_1 tests whether the two thresholds μ_0 and μ_1 differ; that is, whether there is an interval where not contributing to either side predominates. For all PACs, μ_1 is estimated to be positive and quite distinct from zero, indicating a potentially wide range over which noncontribution is the most likely outcome.

The incumbent's rating plays an important, though somewhat complicated role in the contribution decisions of the PACs. For all of the larger PACs (1–6 in Table 1), the probability of making a contribution to the incumbent increases significantly with the rating per se; this is true for only half of the smaller PACs (7–12). The expected closeness of the election reinforces the effect of the rating (i.e., β_2 is significantly greater than 0) in all but two cases (UAW and UMW). A highly rated incumbent is more likely to get contributions if he is in a close race. The challenger of a poorly rated in-

[3] Poole and Romer indicate that, when PACs are aggregated, results of estimating contribution equations are insensitive as to whether margin is defined by the previous or the current election. We use the past election margin, since it is exogenous with respect to current contributions. The 25 percent criterion for a "close race" is a bit more broad than the 20 percent criterion used by Gopoian or, implicitly, by Wright. By our definition, 35.5 percent of the 386 incumbents faced "close elections."

cumbent is more likely to receive a contribution if the contest is expected to be close.

We can use the estimated parameters to get a sense of the impact of being in a close race. For example, consider the NFIB PAC. For incumbents in close races, contributing to the incumbent was the most likely predicted outcome if the incumbent had a rating over 68. For incumbents not in close races ($M = 0$), a positive contribution was the most likely choice only if the rating exceeded 79. Closeness matters at the other end of the spectrum, too. In close races, contribution to the challenger was the most likely choice if the incumbent was rated below 22. In a race that is not expected to be close, an incumbent would have to have a rating under 6 before contribution to the challenger is estimated to be more likely than noncontribution. A similar pattern holds for most of the other PACs, though for the smaller PACs, noncontribution is generally predicted to be the most likely choice, regardless of the closeness of the race.

As a simple measure of overall predictive power, we computed for each PAC the proportionate reduction in error, *PRE*:

$$(4) \qquad PRE \equiv 1 - [\text{classification errors made by probit model}] / [\text{errors made by always predicting modal category}].$$

The modal category is always noncontribution, except for the UAW (in which case, it is contribution to the incumbent). For the six large PACs, each of which made contributions in at least 100 races, the probit model has a *PRE* between .24 and .51. If the classifications for these PACs are pooled, the probit model's *PRE* is .52. These six large contributors made 181 contributions to challengers. Only 5 of these "negative" contributions were predicted "positive" by our model. The same organizations made 861 contributions to incumbents. None of these were predicted by the probit model to have gone to challengers.

On the other hand, for small PACs, the probit model does not do any better than the naive prediction that no contributions will be made. Since noncontribution is estimated as the most likely choice in nearly all cases for these PACs, their *PRE*s are between 0 and .1, even if all the classifications are pooled.[4]

Other specifications in which the latent variable depended on margin per se, on incumbent's seniority, or on a quadratic function of the ratings did no better and, in most instances, did worse than the simple structure whose estimates are reported in Table 1.

IV. Who Gets How Much

Conditional on the incumbent's getting positive or negative money from a PAC, we estimated the amount contributed. We treated this amount as a linear function of independent variables plus a normal error, with separate variances and coefficients for the positive and negative regimes. Because we assume the "how much" decision has no parameters in common with the "who" decision, we can estimate these parameters by *OLS* applied to the positive money and negative money subsamples for each PAC.

We again experimented with a variety of specifications common to the literature, with key variables including ratings, seniority, and margin, as well as interactions among them. The most striking thing about our results is how little of the variation in contributions is explained by the combination of these variables. Only in rare cases did our regressions account for more than 30 percent of the variance. Adjusted r^2 around 0.1 was typical. In many cases, we could not reject (at $p = 0.05$) the hypothesis that the estimated parameters were all equal to zero.

These results are somewhat weak compared to those reported by Gopoian. The major difference is likely due to the fact that each of his regressions includes positive money, zeros, and negative money. Estimates based on such specifications are really picking up the "who gets" effects captured by our probit model rather than the differen-

[4] For the specification in Table 1, we also computed for each PAC the estimated geometric mean probability (*gmp*), which is the log-likelihood divided by N (= 386) and exponentiated. For the large PACs, *gmp* is between .53 and .58; for the small PACs, *gmp* is between .61 and .85.

tial allocation among those actually receiving funds ("how much").

While overall explanatory power is generally poor, we did find the magnitude of positive contributions to be positively related to the PAC's rating of the incumbent in most cases. The effect of electoral margin also tended to echo our "who gets" results. As with our probit results, an incumbent's seniority appears to have no systematic effect on contributions to him.

For the four PACs that made contributions to challengers in more than 30 races, we estimated "how much" regressions for negative money. The results are similar to those for positive money. Once again there is low explanatory power, positive results for margin, and no systematic effect for seniority. In addition, rating appears to relate less well to challenger contributions than to incumbent contributions, perhaps because we have not included evaluations of the challenger in the model.

V. Conclusion

A simple trichotomous probit model does reasonably well in explaining which incumbent congressmen receive support from large PACs in their reelection bids. For large as well as small PACs, the likelihood of a contribution is greater the more highly the PAC rates the incumbent's voting record on issues deemed important by the PAC's parent organization. If the incumbent is expected to be in a close race, the likelihood of both negative (for low-rated incumbents) and positive (for high-rated incumbents) contributions is increased. The smaller PACs in our sample, however, made contributions to only a few races, so for them our probit estimates yield poor predictions.

Moving from contributions favoring the incumbent, to "sitting out" the race, to contributions for the challenger is responsive to the rating of the incumbent. Our results easily differentiate, in terms of ratings, contributions to incumbents from contributions to challengers. All the same, the ratings alone do not allow us to pick out races that attract the attention of a PAC from those that do not. Similarly, we can account only very partially for the variations in the contributions given to those candidates who actually got money from a PAC.

These findings suggest an open research agenda. One approach would be to add additional variables, such as relevant committee assignments or geographical indices (for example, in the case of the UAW, being a member from Michigan), as have been done in earlier studies. Such an approach would marginally improve predictive power, though it has not universally done so when tried in other analyses. A more enticing alternative would be to recognize that the relationship of a representative to a PAC is a dynamic one. It is possible that, when we look at a panel of elections, we will find that individuals who get positive contributions from a PAC tend to become repeat customers (much like an academic who gets his foot in a foundation's door). If so, we could point to a long-term client relationship between a PAC and the relatively few politicians it chooses to support.

REFERENCES

Gopoian, J. David, "What Makes PACs Tick? Analysis of the Allocation Patterns of Economic Interest Groups," *American Journal of Political Science*, May 1984, *28*, 259–81.

Jacobson, Gary C., "Money and Votes Reconsidered: Congressional Elections, 1972–1982," *Public Choice*, No. 1, 1985, *47*, 7–62.

McKelvey, Richard D. and Zavoina, William, "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology*, 1975, *4*, 103–20.

Poole, Keith T. and Romer, Thomas, "Patterns of Political Action Committee Contributions to the 1980 Campaigns for the U.S. House of Representatives," *Public Choice*, No. 1, 1985, *47*, 63–111.

Wright, John R., "PACs, Contributions, and Roll Calls: An Organizational Perspective," *American Political Science Review*, June 1985, *79*, 400–14.

Bargaining and Agenda Formation in Legislatures

By DAVID P. BARON AND JOHN FEREJOHN *

A legislature is a collection of members choosing among a set of alternatives according to some voting rule. Researchers following Kenneth Arrow (1963) have established that except in special cases, no voting equilibrium exists in this setting, that is, there is no alternative that defeats every other alternative. Moreover, when no voting equilibrium exists, any alternative y may arise as the outcome of the proposal-making and voting process: that is, starting at any alternative x it is possible to construct a sequence of alternatives such that each one defeats its predecessor and leads from x to y. (See Richard McKelvey, 1976.)

Traditionally, these results have been given two interpretations. William Riker (1982), for example, has argued that they imply that the outcomes of voting processes are entirely unpredictable and that, as a result, nothing can be said about the likely outcomes except to remark on their arbitrariness. Others have argued that the absence of a majority rule equilibrium allows a member in the position of an "agenda setter" to guide the voting process to attain any alternative he prefers.

When an agenda is exogenously imposed, however, members will take account of its structure and vote in a sophisticated manner, which, as demonstrated by Robin Farquharson (1969), and Nicholas Miller (1980) results in a unique equilibrium outcomes. Even though "admissible" agendas—agendas that can be formed under the rules of the legislature—always produce unique outcomes, the set of admissible agendas, and therefore of outcomes, is very large. In this sense, classical agenda theory does not possess sharp predictive power for the behavior of voting institutions.

In real legislatures, agendas are not imposed by an external party but are built instead by members of the voting body itself. Agendas are thus viewed as endogenously formed by the members of a voting body who make substantive and procedural motions under prespecified rules. Thus, one would expect the outcome of the voting process to reflect the institutional structure of the proposal generation process: the rules that specify which members are able to put motions on the floor, which amendments if any can be made, and how and when motions may be brought to a vote.

From this point of view, the structure of rules governing agenda formation are crucially important. The purpose of the present paper is to develop a noncooperative theory of majoritarian legislatures that permits the examination of these rules. In particular we focus on two commonly observed rules for proposal generation: the closed rule, in which proposals may not be amended, and the open rule, which permits amendments to motions.

We show that majority voting procedures with endogenous agendas generally produce "essentially unique" outcomes. Moreover, these outcomes are *ex post* asymmetric under either an open rule or closed rules, and that an advantage falls to whomever is recognized first. Finally, we show that, in contrast to models of bilateral bargaining under complete information, proposals are not necessarily accepted immediately.

I. The Model of a Legislature

A. *Legislative Structure*

For analytical convenience we assume that the task before the legislature is to divide a dollar among its members according to

* Graduate School of Business, and Department of Political Science and Hoover Institution, respectively, Stanford University, Stanford, CA 94305. This research has been supported by NSF grant nos. SES-8310597 and IST-8606157.

majority rule. Each member is assumed to have selfish, risk-neutral preferences, representable by a von Neumann-Morgenstern utility function. Preferences and the legislative rules are assumed to be common knowledge, so the model involves full information. Although this model is stylized, the division problem is one in which the members have deeply conflicting preferences, and in which there is no majority rule equilibrium in the standard social choice framework. To simplify the presentation, the legislature will be assumed to have three members.

The legislature is governed by a sequential recognition rule in which members compete for recognition to make a proposal. It will turn out that recognition is valuable in equilibrium and so every member will attempt to be recognized. Therefore, the legislature will have to resort to some rule for deciding who shall have the floor, and since side-payments are prohibited, we assume that the legislature adopts a random recognition rule.

Thus at the beginning of a session, member i has a probability p_i of being recognized, and if recognized by the chair, he may propose a bill that specifies how the dollar is to be divided. This is then the motion on the floor, and the resolution of that motion is governed by the rules of the legislature. Under a *closed rule* the motion is voted on immediately, and if approved the legislative adjourns. If the motion fails, the next session begins with a member being recognized to propose another bill. Under an *open rule*, after a bill has been proposed, another member j is recognized (with probability $p_j/\Sigma_{k \neq i} p_k$,) and he or she may either offer an amendment or move the previous question.[1] If the previous question is moved, the legislature votes on whether or not to accept the proposed division of the dollar. If an amendment is offered by member j, it becomes the question on the floor, and another member l may be recognized (with probability $p_l/\Sigma_{k \neq j} p_k$). If the previous question is moved at this point, the amendment is put to a vote and, if it wins, the bill as amended becomes the new motion on the floor. If the amendment fails, the original bill becomes the question on the floor. The process continues with members proposing amendments or moving the previous question and forcing a vote until the bill itself, as amended, is put to a vote. If the bill is agreed to, the process stops; otherwise, another member is recognized to propose a new bill. Whenever the previous question is moved, voting takes place sequentially and openly. That is, there is a fixed order in which each member must announce his vote on the question before the body and every other member may observe each vote as it is cast. Members thus have full information about previous votes.[2]

Unlike the two-member bargaining problem studied by Ariel Rubinstein (1982), agents need not exhibit impatience to be motivated to approve a bill. If a member fails to vote for a bill, that member runs the risk that others will exclude him or her from the division of the dollar in the next session. In order to increase the generality of our results, however, we assume that agents are impatient and have a common discount factor $\delta \leq 1$. At times, the limit of the division as δ converges to one from below will be studied, which corresponds to the case in which the period between offers (which we call a session) is very short.

In contrast to models in which the agenda is exogenous and hence all members have perfect information about the sequence of votes to be taken, members in the legislature considered here must form expectations about future proposals, motions, and votes. The process of proposal generation and voting yields an extensive-form game with an infinite game tree. A strategy in this game is a prescription of what motion to make at each point at which the member is recognized, and a prescription of how to vote whenever a vote is required.

[1] See Ferejohn, Morris Fiorina, and McKelvey (1986) for a similar specification.

[2] If voting were by secret ballot, the set of subgame perfect equilibria would be very large, and a stronger equilibrium concept such as that of sophisticated equilibrium would be required. See Farquharson and Hervé Moulin (1979).

The game is generated as follows: The chair randomly recognizes a member to make a proposal. This member then may propose a division of the dollar, which will be called a *bill.* The chair then recognizes another member who may either propose an alternative division or require the members to vote on the motion on the floor. If the second member proposes an alternative division, it is regarded as an *amendment* to the bill. The chair then recognizes another member, who is distinct from the second member but not necessarily from the first, who may either propose another division (an amendment to the amendment) or require the members to vote on the previous motion; the amendment. The chair then recognizes another member and so on. The game terminates when a bill, as amended, passes.

A *history*, h_t, of the game up until session t is a specification of who had a move at each previous session and the move selected by each member at every time he had a move to make. If H_t denotes the set of histories, then a *strategy* for member i is a sequence of functions s_t mapping H_t into his or her available actions at t. An important feature of this formulation is that at any time that an agent is to take an action, he or she knows which history has occurred, so the game is one of perfect information.

Members are not able to make binding commitments to vote in a particular manner or to offer a particular proposal. Thus, an equilibrium strategy must be "self-enforcing" in the sense that the member would wish to execute it at each point in the game tree at which he or she has an opportunity to act. Therefore, attention is restricted to subgame perfect equilibria. An equilibrium collection of strategies is subgame perfect if the restriction of those strategies to any subgame constitutes an equilibrium in that subgame. Because votes and proposals occur sequentially and openly, subgame perfect equilibria correspond to dominant solvable solutions (see Moulin). For any particular subgame perfect equilibrium, the *continuation value* of a subgame is defined as the vector of values to the members resulting from the play of that subgame perfect equilibrium strategies.

B. *An Illustration*: *A Closed Rule with a Finite Number of Sessions*

To illustrate the basic structure of the model, consider the simple case of a three-member legislature governed by majority rule, with equal probabilities of recognition, and in which there is a closed rule that prohibits amendments once a bill has been offered. The agenda will be assumed to be a finite with at most two proposals made prior to adjournment. With a closed rule; the member recognized in the first session may offer a bill specifying how the dollar is to be divided, which must then be voted up or down. If this bill receives a majority vote, the dollar is divided according to the bill and the legislature adjourns. If it is defeated, the second session commences and a member is selected at random to offer a bill. If there is no agreement in the second session, the legislature dissolves, and each member receives zero.

In this case, the subgame perfect equilibria are easily characterized. Note that, if the first offer is rejected and a second (and last) session were to take place, whoever is recognized will propose to take the whole dollar and this proposal will be accepted by the other members who stand to get zero in any case. Thus, if each member has an equal likelihood of gaining recognition in the second session, the continuation value of the game prior to anyone gaining recognition in the second stage must be equal to $1/3$ for each member. Thus, in the first session whoever is recognized can propose to offer $\delta/3$ to one other player and keep $1-\delta/3$ for himself. This proposal will be accepted and the legislature will adjourn immediately.

The important features of this equilibrium are: A) *ex post*, 1) the allocation reflects the majoritarian distribution of power in that only a minimal majority of members receives a positive payoff, and 2) the member recognized in the first session has agenda power and thus receives the largest allocation; B) *ex ante* the symmetry of the legislature is reflected in the fact that every member attaches the same value to the game. Finally, as in the bargaining models of Rubinstein and Kenneth Binmore (1986), the initial offer is accepted and the legislature adjourns after only one bill has been proposed and voted on. This last feature of the equilibrium is not due to impatience but results from the probability that the member will not be recognized in the last session.

The majoritarian equilibrium exhibited here involves the division of the dollar among a minimal majority of the legislature. This minimal majority is not a coalition in the sense that that term is used in cooperative game theory and in the social choice literature. The members of the majority in the legislature considered here act noncooperatively and each finds it in his own interests to act as specified in the equilibrium.

II. A Folk Theorem under a Closed Rule

While the case considered in the previous section has a unique subgame perfect equilibrium, this result does not extend to legislatures that have no limitation on the number of sessions. When the number of sessions is unlimited, any division of the dollar may be supported as a subgame perfect equilibrium if the members are not too impatient. In view of the uniqueness results for two-member bargaining, it is perhaps surprising that multimember bargaining yields complete indeterminacy.[3]

Remark: For the sake of simplicity, this result is stated for a majority-rule legislature but it should be clear from the nature of the argument that such a theorem holds for any voting rule as long as the legislature contains at least three members.

PROPOSITION 1: *For an n member, majority-rule legislature with an infinite number of sessions and a closed rule, if* $\delta > (n+1)/2(n-1)$ *and* $n > 4$, *then any division of the dollar may be supported as a subgame perfect equilibrium. In every equilibrium the first offer is accepted.*[4]

The idea of the proof is simple: In order to support an arbitrary division x, a strategy configuration is constructed in such a way that any member recognized is certain to be punished if he deviates by failing to propose the prescribed division x, when required to do so, or by failing to punish someone who deviated before him.

The punishment scheme required for the folk theorem must be "infinitely nested" to allow any alternative to be supported as a subgame perfect equilibrium. Thus, the members of the legislature are able to base their choice at any stage of the game on the whole history of play to that point. If such history dependence is not allowed, the set of alternatives that may be supported is reduced. One way to achieve this restriction is to require members to choose among stationary strategies.

Definition: A strategy is *stationary* if it dictates that a member take the same action in identical subgames.

For example, if, in two different sessions, a member is recognized and there are no motions on the floor, he must make the same proposal in both sessions. Clearly, the strategies required for the folk theorem fail to satisfy this restriction, since what a member is required to propose depends on the history of play leading to the subgame.

III. A Closed Rule and Stationary Strategies in an Infinite Session Legislature

We now provide a general treatment of a legislature with an infinite number of periods in which proposals are considered under a closed rule. A proposal made in a session is then either approved by a majority and the legislature adjourns or it is rejected, and the next session commences with a member recognized at random by the chair. Without sacrifice of generality, we restrict our investigation to the three-member case. Member i has a probability p_i of being recognized and each member employs a common discount factor $\delta \leq 1$. For convenience attention is restricted to the symmetric case in which p_i is equal to 1/3. The following proposition indicates that a majoritarian outcome results with stationary strategies.

[3] This observation has been made by Maria Herrero (1985) who considered bargaining under a unanimity rule and a fixed order of recognition.

[4] The construction in this proof is essentially the same as that found in Herrero, and is presented in our earlier paper (1986).

PROPOSITION 2: *For all $0 < \delta \le 1$, a strategy configuration is a stationary subgame perfect equilibrium in an infinite-session, majority-rule legislature governed by a closed rule if and only if it has the following form*:

1. *Each member recognized proposes to receive $1 - \delta/3$ of the dollar and offers one other member $\delta/3$ of the dollar.*
2. *Each member votes for any proposal in which he receives at least $\delta/3$.*
3. *The strategy configurations are balanced in the sense that each member receives an offer of $\delta/3$ from one other member in each session, for example, from the member on his left.*

Proposition 2 provides a rationale for the restriction to stationary strategies when the agenda is governed by a closed rule. The division of the dollar specified in the proposition is the limit of the divisions chosen in finite session games as the number of sessions increases.

Proposition 2 may be thought of as the natural extension of the Rubinstein model to a three-member legislature. As in two-member bargaining theory, the first member recognized proposes a bill that is sufficiently attractive to a majority that it is immediately accepted and the legislature adjourns after one motion. As in the model with finitely many sessions, however, this is due to majority rule rather than impatience and thus holds for all $\delta < 1$.

Unlike the bilateral case Rubinstein studied, division under majority rule is *ex post* asymmetric even if $\delta = 1$. Since a majority can exclude a minority and divide the dollar among themselves, the recognized member has an incentive to exclude as large a minority as possible. A minimal winning majority will thus be formed with the dollar divided among the members of that majority. To determine how the dollar will be divided, consider the case in which $\delta = 1$. Equal division among all members involves 1/3 for each, so the one-third share of the excluded member is to be allocated between the member recognized and the other member of the winning majority. Under majority rule that member recognized is able to capture the entire share of the excluded member. With impatience that member also captures a premium of $(1 - \delta)/3$ due to the impatience of his partner.

The condition in item 3 of Proposition 2 is a balance property of the stationary strategy equilibrium. If the strategies were not balanced so that each member receives exactly one offer to be in the majority, then the disadvantaged member would be a preferred "partner" in the previous session and would receive offers from each of the other members if they were recognized. This would then bid up his continuation value until he is no longer a preferred partner. Competition to be a preferred partner would then equalize the continuation values.

Remark: In the present case, the requirement that members restrict themselves to stationary strategies implies the balance property of the equilibrium outcome in Proposition 2. For example, if the continuation values of the game after the first session resulting from some subgame perfect equilibrium are $v_i = 1/3$, $i = 1,2,3$, then proposals x^i,

$$\begin{aligned}(x^1 &= (1-\delta/3, \delta/3, 0),\\ x^2 &= (0, 1-\delta/3, \delta/3),\\ x^3 &= (0, \delta/3, 1-\delta/3)),\end{aligned}$$

in session one and thereafter the strategies yielding the continuation value form an equilibrium. This equilibrium also has the properties that the first bill proposed is accepted and that the member recognized receives $1 - \delta/3$, one other member receives $\delta/3$, and the other receives zero. Thus, in one sense, the equilibrium is *ex post* majoritarian.

Remark: While Proposition 2 is established in the special case of a three-member, majority-rule legislature, extensions to the n member, majority-rule legislature are immediate: The initial proposer must offer the continuation value δ/n to $(n-1)/2$ members (assuming that n is odd) and keep $1 - \delta(n-1)/2n$ for himself, and, as before, each member must receive the same number of offers in a stationary equilibrium. As n increases, the share of the member recognized approaches $1-\delta/2$, so the value of recognition is decreasing in the size of the legislature. For the case of k member majority rule in which k members are required to adopt a proposal, the member recognized must again offer δ/n to $k-1$ members and keep $1-\delta(k-1)/n$ for himself. Note that as k goes to n, the share retained by the initial proposer goes to $1-\delta(n-1)/n$, which for $\delta=1$ equals $1/n$. This is the unanimity allocation under stationary strategies.[5] As k goes to one, the member recognized receives the entire dollar. This may be thought of as awarding a property right to the member recognized.

IV. Amendments: A Simple Open Rule

With an open rule, motions on the floor may be subject to amendment. An amendment is itself an allocation of the dollar and may be viewed as a substitute for the motion on the floor. The simplest open rule is one in which no more than one amendment may be on the floor at any time. That is, once an amendment is offered it must be disposed of by bringing it to an immediate vote. If the amendment fails, the prior motion remains on the floor, and the next session commences. If the amendment obtains a majority, the amendment becomes the motion on the floor, and the next session commences. At this point another amendment may be offered or the previous question on the bill may be moved.[6] This rule is properly termed an open rule because there is no limit to the number of amendments that may be offered before the bill itself is brought to a vote. In this setting, discounting is assumed to occur whenever a new amendment is moved.

The legislature operating under this open rule may thus be described as follows. Each member has a 1/3 probability of gaining initial recognition. A member recognized may offer a bill. Then, each of the other two members has a 1/2 probability of being recognized for the purpose of making an amendment or moving the previous question. If the previous question is moved, the bill is voted and if approved, the dollar is divided. If the bill is defeated, one of the two members who did not move the question is recognized with equal probability. If an amendment is offered, it must be voted against the bill before another amendment or moving the previous question is in order. The winner then becomes the motion on the floor, and each member other than the amender has a probability of 1/2 of recognition for the purpose of offering an amendment or moving the previous question. The process continues in this fashion until the previous question is moved on a bill and the bill passes.

In this open rule procedure the power of the member recognized is limited because no member will ever be in a position to introduce a bill or an amendment that cannot itself be subjected to an amendment. That is, after an amendment is moved and voted on, another member will be recognized who may introduce an amendment, if that is in order, or may move the question on the bill, which allows yet another member to offer an amendment. Thus, whoever proposes a bill or an amendment must take account of the fact that other members will subsequently be recognized. Another difference between an open rule and a closed rule is that a vote takes place only after a member has been recognized. To simplify the mathematics, a member who is indifferent between two proposals will be assumed to vote for the one proposed last.[7] A stationary equilibrium in a legislature operating under a simple open rule is characterized in the following proposition.

PROPOSITION 3: *In a legislature with a simple open rule, equal probabilities of recognition, and δ near one, a stationary subgame perfect equilibrium strategy configuration has the following form*:

1. *The member recognized first offers $1-y_1$ to another member and proposes to keep y_1 for himself, where $y_1=(1-\delta/4)/(1+\delta/4)$.*

[5] B. Dutta and Louis Gevers (1981) obtained this result.

[6] In the terminology of Ferejohn et al., such amendment rules exhibit a "depth" limitation. In legislative parlance, no second-degree amendments may be offered.

[7] Otherwise, the set of amendments that defeat the proposal on the floor is open, and the equilibrium concept has to be weakened to that of an ε equilibrium.

2. *If the next member recognized, has been offered at least* $1 - y_1$, *she moves the previous question and the question is approved; otherwise she proposes to offer* $1 - y_1$ *to the member that did not make the previous motion and to keep* y_1 *for herself.*[8]

The following are the important features of the stationary strategy equilibria for this open rule when δ is near one. First, *ex post*, the dollar is divided among a minimum winning coalition. Second, and unlike the usual case in bargaining models with complete information, the initial offer is not necessarily accepted. Rather, with some probability, an amendment is offered and the legislature continues to the next session. When impatience is low, the probability is thus positive that the legislature will not reach an agreement in any finite number of sessions. Thus, if $\delta < 1$, the sum of the continuation values is less than one. Third, the power of the member initially recognized is greater the greater is the impatience of the members. Fourth, while amendments may be offered in equilibrium, the initial proposer is still advantaged. Thus, as in the case of the closed rule, recognition is valuable. Fifth, recognition is not as valuable as under a closed rule. For example, for $\delta = 1$, $y_1 = 3/5$ and so the value of the game to the first member recognized is 2/5.

V. Conclusions

Our purpose in this paper has been to introduce a game-theoretic model that permits the study of agenda formation in legislatures. To do this we made a number of simplifying assumptions that deserve close examination in future work. The principal result of this paper is that, in a complete information setting, legislative outcomes with endogenous agenda formation are quite determinate. This determinacy follows from the bargaining structure of the agenda formation process and is due largely to the ability of members to make proposals. The actual payoffs depend also on the voting rule as well as on the rules of agenda formation. These findings stand in sharp contrast to the results of agenda voting models that do not allow for endogenous agenda formation.

REFERENCES

Arrow, Kenneth, *Social Choice and Individual Values*, New Haven: Yale University Press, 1963.

Baron, David and Ferejohn, John, "Bargaining in Legislatures," unpublished, Stanford University, 1986.

Binmore, Kenneth, "Bargaining and Coalitions," in Alvin Roth, *Game-Theoretic Models of Bargaining*, Cambridge: Cambridge University Press, 1986.

Dutta, B. and Gevers, Louis, "On Majority Rules and Perfect Equilibrium Allocation of a Shrinking Cake," unpublished, Namur University, 1981.

Farquharson, Robin, *Theory of Voting*, New Haven: Yale University Press, 1969.

Ferejohn, John, Fiorina, Morris and McKelvey, Richard, "Sophisticated Voting and Agenda Independence in the Distributive Politics Setting," *American Journal of Political Science*, forthcoming 1987.

Herrero, Maria, "A Strategic Bargaining Approach to Market Institutions," unpublished doctoral dissertation, London University, 1985.

McKelvey, Richard, "Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control," *Journal of Economic Theory*, June 1976, *12*, 472–82.

Miller, Nicholas, "A New Solution Set for Tournaments and Majority Voting," *American Journal of Political Science*, February 1980, *24*, 68–96.

Moulin, Hervé, "Dominance Solvable Voting Schemes," *Econometrica*, November 1979, *47*, 1337–51.

Riker, William, *Liberalism Against Populism*, San Francisco: Freeman, 1982.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *50*, 97–109.

[8] The proof is provided in our earlier paper.

THE ECONOMICS OF DISCRIMINATION THIRTY YEARS LATER[†]

The Conceit of Labor Market Discrimination

By THOMAS F. D'AMICO*

"In view of the importance of discrimination, it may seem surprising that economists have neglected its study" (2nd ed. 1971, p. 10). That assertion, made by Gary Becker in 1957, continues to ring true for many even today. In 1957 the very small amount of published matter on the subject was tangible evidence of neglect by the profession. Over the last three decades, however, that literature, virtually all of it heavily influenced by Becker's work, has burgeoned. And yet neglect, albeit of a different sort, is still evident.

The neoclassical theory of discrimination, introduced by Becker and later refined and extended by Kenneth Arrow (1972; 1973) among others, has the dubious distinction of being, far and away, the most influential economic theory of discriminatory income differentials and, simultaneously, the most heavily attacked. This paper provides a brief survey of the principal tenets of that theory and of the variety of criticisms that have been leveled at it.

I. The Costs of Discrimination

Discrimination is generally understood to exist when some superficial characteristic (skin pigmentation, for example) is used in an attempt to restrict individuals' access to the available economic, political, and social opportunities for advancement. "Superficial," in this context, signifies that the characteristic being used for discriminatory purposes is either largely or completely unrelated to the individuals' actual or potential talent, skills, and drive. Discrimination is effective when society's tangible and intangible compensations are, in fact, consistently distributed at least in part on the basis of this characteristic, that is, without full regard for the relative productivities of its members.

It follows logically from this that effective discrimination will be costly to society, since it results in a clear and potentially serious loss of efficiency. When society's rewards and penalties are distributed to its members in a manner not consonant with their relative productivities, then at least some scarce resources are bound to be overallocated to relatively unproductive members of the "favored" race (assuming race is the trigger) and underallocated to more-productive members of the race being discriminated against (the "target" population). Society's aggregate real output, therefore, will fall below its potential and the size of this shortfall will depend upon, among other things, the intensity of the discriminatory feelings, the ease with which members of the target population can be identified, and the size of the target population.

It also follows that in a democratic society, racial discrimination, to be more specific, cannot be effective unless there is at least implicit acceptance of the propriety of the distribution-by-race scheme. However, since racial discrimination generates efficiency losses which depend, as has already been suggested, on the size of the target population, the strength of voters' convictions regarding the propriety of discrimination will have to vary directly with the size of the target population, or else effective discrimination will be very difficult to achieve and maintain.

[†] *Discussants*: Gary S. Becker, University of Chicago; Isabel V. Sawhill, The Urban Institute.

*Department of Economics, St. Peter's College, Jersey City, NJ 07306.

This line of thought is consistent with Becker's assertion that "the relative number of *N* [the target population] in the society at large also may be important..." as a determinant of the magnitude of the society's tastes for discrimination and the effectiveness of discrimination in general (1971, pp. 16–18), and that "...a necessary condition for effective discrimination against *N* is that *N* be an economic minority; a sufficient condition is that *N* be a numerical minority..." (p. 27). Whether or not tastes for discrimination are more pronounced, and discrimination itself is more effective as the relative size of the targeted minority grows, however, remains an open question since the targeted minority's political, social, and cultural influences are also likely to grow with its numerical strength.

It should also be noted that this question of the relationship between tastes for discrimination and the effectiveness of discrimination implicitly assumes that the specific mechanisms by which discrimination is effectively pursued are fairly overt and, therefore, able to be spotlighted and rationally evaluated. The more subtle the mechanism, however, the more likely it is to persist unchallenged. If, for example, as the result of a long history of intensely held prejudices, the discrimination-by-race mechanism comes to be subtly and intricately woven into the institutional and cultural fabric of the society, then the intensity of currently held discriminatory tastes may be of little importance as a determinant of the effectiveness of discrimination and, perhaps, may even be completely irrelevant. In a society where discrimination is evidenced in a mixture of covert and open devices, the nature and significance of current, individual tastes for discrimination is only one of a variety of factors which will influence effectiveness; the numerous arenas in which people act collectively—governmental and cultural, for example—are bound to be equally influential.

Aside from society's overall loss of efficiency, effective discrimination also is expected to impose additional costs on the individual members of the target population. Deprived of full participation in the society's production processes, or of full compensation for that participation, members of the target group will, in general, enjoy a lower quality of life relative to that which would have existed in the absence of discrimination and relative, as well, to their peers in the favored race. These personal losses include, but are not limited to, lower per capita real income, poorer living conditions, and lower social status. Paul Baran and Paul Sweezy (1966), Michael Reich (1977), and others have also suggested that under certain circumstances—if, for example, the bargaining power of workers in general is fragmented and weakened by racial tensions—individual working-class members of the favored race may suffer personal earnings losses as well, to the benefit of employers. Becker, on the other hand, contends that "Although the aggregate net incomes of *W* [the discriminators] and *N* are reduced by discrimination, all factors are not affected in the same way: the return to *W* capital and *N* labor decreases, but the return to *W* labor and *N* capital actually increases" (1971, p. 21).

The empirical work on these important differences of opinion is rather shallow but, unfortunately, not unusually so. Indeed, a major theme among critics is the lack of attempts at empirical verification of at least the major implications of the theories and "...the disappointing yield of most hypothesis testing" (Glen Cain, 1985, p. 64).

II. Types of Discrimination

The literature differentiates between two types of discriminatory actions: those occurring within the context of the labor market and those taking place beyond its boundaries. Following Becker's lead, and assuming a society in which whites discriminate against blacks, labor market discrimination exists when white employers or employees have a distaste for association with blacks and conduct their labor market transactions in a way that is intended to minimize or eliminate such contact. According to Becker, these discriminators "...must, in fact, either pay or forfeit income for this privilege" (1971, p. 14). Nonlabor market discrimination, on the other hand, refers to the differential treatment accorded blacks outside of the labor markets, largely before they enter the labor force, and becomes manifest as a confluence of school, home, and neighborhood deficiencies which cause blacks, on average, to be less well prepared and, therefore, less marketable than whites.

While the distinction between labor market and other forms of discrimination seems intuitively appealing, the two forms are really quite difficult if not impossible to separate analytically. Our view of labor market discrimination is never unobstructed and, in operation, the dichotomy turns out to be artificial and misleading.

Although discrimination within the labor market and beyond its boundaries are fundamentally different processes, the outcomes of each are essentially the same or, at least, similar enough to cause some serious theoretical and empirical difficulties. Any earnings differential—between blacks and whites in the United States, for example—is capable of decomposition into a wage gap and a productivity gap. The productivity gap is that portion of the earnings differential attributable to the fact that black workers, on average, have smaller human capital endowments than their white peers. The remaining earnings difference, the wage gap, is attributable to the fact that black human capital yields a smaller unit return than that of whites. The size of this wage gap is typically taken as an index of the virility of labor market discrimination while the literature is silent, generally, on the nature and causes of the productivity gap. If, however, one assumes that such attributes as innate talents, tastes, and drive are randomly distributed throughout the population without regard to race, then it is difficult to avoid the conclusion that this productivity gap is, itself, also emanating from discrimination, largely but not wholly nonlabor market in origin.

The problem, however, is considerably more complex than this simple empirical technique lets on. Part of the productivity gap can, in fact, be traced to the presence of labor market discrimination, and should therefore be included in any measure of the power of that phenomenon. Since the acquisition of human capital is ordinarily a costly process undertaken, in large part with the expectation of a tangible future payoff, it is not at all implausible that an individual's willingness and ability to acquire human capital will be significantly influenced by the size of those expected payoffs. Labor market discrimination, then, by directly diminishing the relative wages of blacks also serves to dampen the acquisition of human capital by blacks, thus lowering their earnings further (but through the productivity gap).

Moreover, because of chronic misspecification problems it is highly probable that some of the group differences in productivity-related characteristics attributable to nonlabor market discrimination are missed by the productivity gap estimates and mistakenly subsumed, instead, by the residual—the wage gap. Equally troublesome is the possibility that some part of the lower relative return to black human capital is correctly attributable to nonlabor market forms of discriminatory behavior and correctly subsumed by the residual wage gap. For example, statistical theories of discrimination predict that employers, if they perceive blacks as being generally less productive than whites and if it is difficult to measure the actual productivity of an individual applicant, will use race as a convenient screening device and make lower wage offers to blacks than whites. In this case, of course, blacks with above-average productivity will receive below-average returns (relative to whites). One can argue that these lower returns, to some extent, reflect the presence of nonlabor market discrimination, which contributed to the unfavorable classifications to begin with.

While the strength of these effects is simply not know, their existence is difficult to dismiss a priori. The main point here, however, is that the wage gap is likely to measure both more and less than originally intended. It misses some of the earnings losses subsumed by the productivity gap which should, instead, be attributed to labor market discrimination and it arbitrarily includes some of the impact of nonlabor market discrimination. Serious questions exist, therefore, about the reigning empirical devices used to isolate the effects of labor market discrimination and, in turn, about the theoretical models upon which they rely and which generally are not alive to these possibilities.

Although it is not clear that economic analysis is any more suited to the study of labor market than to nonlabor market discrimination and no reliable evidence exists that labor market discrimination is significantly more prevalent or costly, labor market discrimination has received and continues to receive the lion's share of attention in the theoretical and empirical literature. This in spite of the important linkages and feedbacks between the two forms. Becker, in part, seems to agree that our study of discrimination phenomena has been too narrowly drawn: "...relatively little is known about [the] quantitative importance and socioeconomic determinants [of discrimination].... Considerably more study of institutional arrangements is required in order to know more about the factors determining the talents that minorities are permitted to bring to the market place" (1972, p. 210).

The labor market does not stand alone, insulated from the society at large. Rather, these markets are nestled into society, processing and filtering the prevailing customs and political and social arrangements and translating these into particular, tangible economic outcomes. These outcomes, therefore, reflect more than just the organizational and technical characteristics of production and the resource endowments of the individual and society. Market outcomes reflect social relations and feed back on them, and nowhere is the importance of this interaction more pronounced than in the dynamics of discrimination.

III. Long-Run Instability

In general, Becker's model attempts to explain racial wage differentials among homogeneous units of labor in terms of the deviant market behavior of individual, noncollusive employers, employees, and consumers. The behavior of white employers, for example, is deviant in the sense that profit maximization is not the primary motivation. Instead, white employers are assumed to have a "taste" for discrimination (the origins of which are important but unstipulated) and they are assumed to be willing to forego some profit to avoid the "psychic costs" of interracial contact. Arrow (1972) and Ray Marshall (1974), among others, have argued that discriminators seek to avoid social rather than physical contact, but Cain points out that while a restatement in these terms would seem to yield a more realistic model more clearly applicable to the types of sex and race discrimination actually observed, "these interpretations...do not negate Becker's central point, which was the establishment of an equilibrium differential in favor of white workers" (p. 35). Criticism of a different sort is raised by Finis Welch, who decries the absence of any "...empirical attempt to verify that profits of firms depend upon the racial composition of their work force" (1975, p. 70).

In simplified terms, employers' aversion to contact with black workers lowers the demand for black workers and depresses the relative wages of blacks accordingly. With heterogeneous tastes for discrimination, the extent of the disadvantage for blacks will depend, as Becker indicates, on the average intensity of employers' discriminatory tastes and the dispersion around that average. The competitive advantage available to nondiscriminating employers, however, is quite obvious. In the short run, they will incur lower per unit labor costs by hiring black workers instead of white, and earn higher profits than those employers who are hostile toward blacks. Over the long run, the least discriminatory employers will expand at the expense of those employers most hostile to blacks (i.e., those who exhibit the most deviant market-behavior patterns) and the black-white wage gap will be mitigated. Competitive pressures will serve, in the long run, to restore wage parity between black and white workers who are perfect technical substitutes if there is at least one completely nondiscriminating employer in the pool. Becker's model, therefore, "...predicts the absence of the phenomenon it was designed to explain" (Arrow, 1972, p. 192). Becker anticipated this result but discounted it by arguing that the ability of the least discriminatory firms to expand in the long run will be compromised "...if costs rise sharply with output..." (1971, p. 44).

According to Welch, "In view of the long history of discrimination, it seems to me that explanations which are stable only in the short run are unsatisfactory" (1967, p. 227). The search for a more satisfactory explanation led, during the 1970's, to a proliferation of theories, many of which attempted to modify Becker's model by introducing some market imperfection capable of short-circuiting the long-run competitive impulse toward wage parity. Other, heterodox approaches—institutionalist and racial—have also been developed. Like Becker's model, however, none of these have had the benefit of rigorous empirical testing and verification, the most basic elements in the progression of any scientific inquiry. (These alternative theoretical developments are surveyed in a number of places. See, for example, Cain or my 1983 study.)

IV. Conclusion

Neoclassical economic theories of discrimination are invariably theories of labor market discrimination. As such they tend to be too narrowly drawn, paying very little attention to aspects of the problem, like nonlabor market discrimination and collective action, which are central to our understanding of discrimination in general and labor market discrimination in particular. Moreover, they are typically not etiological in nature, but attend more narrowly only to the direct economic consequences (indeed, the direct demand-oriented economic consequences) of this one particular manifestation of discriminatory behavior. Finally, they are almost purely speculative in nature, the lack of systematic empirical verification emanating in part from the theories themselves, which tend to yield vague and untestable conclusions.

There is no doubt that Becker's pioneering work heightened the profession's interest in an issue that had been, up to that point, badly neglected and that his work served as a catalyst for an enormous amount of economic research. However, the overall results of that research program, if one can judge such a thing after just three decades, have been disappointing. The processes of discrimination, its nature and causes, the width and depth of its debilitating effects on the target population and the appropriate correctives are still not well understood.

In truth, that research program has always seemed to be marked by a certain disarray —fragmented and straining under the weight of the *ad hoc* theorizing and sometimes cavalier data manipulations which have been its hallmark. The base upon which the research was erected appears, in hindsight, to have been too narrow to support the elaborate and complex structure of questions that needed to be answered. A broader theoretical foundation is necessary, one that allows a suitable synthesis of the operative social, political, and economic forces.

REFERENCES

Arrow, Kenneth J., "Models of Job Discrimination" and "Some Mathematical Models of Race in the Labor Market," in A. H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington: Lexington Books, 1972.

______, "The Theory of Discrimination" in O. Ashenfelter and A. Rees, eds., *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973.

Baran, Paul A. and Sweezy, Paul M., *Monopoly Capital*, New York: Monthly Review Press, 1966.

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957; 2d ed., 1971.

______, "Economic Discrimination," in D. L. Sills, ed., *International Encyclopedia of the Social Sciences*, Vol. 4, New York: Macmillan and Free Press, 1968; reprint ed., 1972.

Cain, Glen G., "The Economic Analysis of Labor Market Discrimination: A Survey," Institute for Research on Poverty Special Report No. 37, University of Wisconsin-Madison, August 1984, rev. March 1985.

D'Amico, Thomas F., *The Economics of Market and Non-Market Racial Discrimination*, Ann Arbor: University Microfilms International, 1983.

Marshall, Ray, "The Economics of Racial Discrimination: A Survey," *Journal of Economic Literature*, September 1974, *12*, 849–71.

Reich, Michael, "The Economics of Racism," in D. M. Gordon, ed., *Problems in Political Economy: An Urban Perspective*, 2nd ed., Lexington: D.C. Heath, 1977.

Welch, Finis R., "Labor Market Discrimination: An Interpretation of Income Differences in the Rural South," *Journal of Political Economy*, June 1967, *75*, 225–40.

______, "Human Capital Theory: Education, Discrimination, and Life Cycles," *American Economic Review Proceedings*, May 1975, *65*, 63–73.

Discrimination: Empirical Evidence from the United States

By FRANCINE D. BLAU AND MARIANNE A. FERBER*

The study of discrimination received a major impetus in the 1960's when increasing social attention focused upon race and gender differentials in market outcomes. Gary Becker's *The Economics of Discrimination* (1957) strongly influenced empirical research by providing a definition of wage discrimination and suggesting a specific way in which it might operate. During the following years new theories were developed and refined in an attempt to explain why there appears to be continued discrimination in spite of market forces presumably operating against it. Similarly, a large amount of empirical work has been done to determine whether and how much discrimination actually exists, and to a lesser extent to test the implications of the various theories. Even so, the hope expressed by Becker in the preface of the second edition (1971) that our understanding of discrimination would increase so rapidly that the materials in his book would become obsolete before another decade began has clearly not been fulfilled. Here, we review what has been learned in the intervening years and suggest some fruitful directions for future research.[1] The focus of this paper, like that of most of the empirical research in this area, is on determining the extent of discrimination rather than on testing alternative models of discrimination.

I. Evidence from Econometric Studies

Becker defined the market discrimination coefficient as the difference between the ratio of two groups' wage rates with and without discrimination. Since it is assumed that, in the absence of discrimination, the members of each group would be paid in accordance with their productivities, wage discrimination may also be defined as pay differences between two groups that are not accounted for by productivity differences. Such a definition of discrimination implies, however, that any wage differences associated with productivity differences which are themselves caused by discrimination should be included in the discrimination coefficient.

In the late 1960's and early 1970's, a widely used method of empirically measuring discrimination was developed. In this approach, separate wage or earnings regressions are estimated for the members of each group, say blacks and whites. The portion of the pay differential due to differences in the returns to a given set of characteristics (i.e., differences in coefficients, including the constant term) is then used as an estimate of discrimination (for further explanation, see Alan Blinder, 1973; Glen Cain). Specifically, let U_r = the unadjusted wage ratio; A_r = the adjusted wage ratio of black to white wages if the two groups had equal values of the explanatory variables; and U_x = the unexplained portion of the wage differential or the estimate of discrimination, then

$$(1)\quad U_r=\sum_i B_{ib}\bar{X}_{ib}/\sum_i B_{iw}\bar{X}_{iw}$$

$$(2)\quad A_r=\sum_i B_{ib}\bar{X}_{iw}/\sum_i B_{iw}\bar{X}_{iw}$$

$$(3)\quad U_x=(1-A_r)/(1-U_r) = \left[\sum_i \bar{X}_{iw}(B_{iw}-B_{ib})\right] \Big/ \left(\sum_i B_{iw}\bar{X}_{iw}-\sum_i B_{ib}\bar{X}_{ib}\right)$$

where B_i and $\bar{X}_i$ are, respectively, the coefficient and mean of the ith variable in an OLS

*Professor of Economics and Labor and Industrial Relations, and Professor of Economics, respectively, University of Illinois-Urbana, Champaign, IL 61820. We are grateful for the comments and suggestions of Barbara Bergmann, Charles Brown, Barry Chiswick, Claudia Goldin, Daniel Hamermesh, Lawrence Kahn, Barbara Reskin, and Robin Williams.

[1] For more detailed reviews, see Blau (1984), Glen Cain (forthcoming), William Darity (1982), and Janice Madden (1985).

wage (earnings) regression and the subscripts b and w denote blacks and whites, respectively. It should be noted that (3) is essentially an index number formulation. The estimate of discrimination will depend upon the weights chosen and the use of any particular set of weights, such as the $\bar{X}_{iw}$ above, is arbitrary.

In evaluating the findings of such studies, it is important to bear in mind that, apart from the index number problem, (3) may yield either an over- or an underestimate of the true extent of labor market discrimination. On the one hand, conventional data sources do not allow for the measurement of all productivity-related characteristics. The concern is generally that if, on average, whites (males) are more highly qualified with respect to these omitted variables, labor market discrimination will be overestimated.[2] On the other hand, differences with respect to some productivity-related characteristics may reflect the indirect effects of discrimination (Blinder). For example, employers may discriminate against blacks (women) in access to training programs or in hiring for particular occupations. Further, if one takes into account the endogeneity of choice variables, we see that labor market discrimination may adversely affect the human capital investment decisions and labor market behavior of groups experiencing discrimination. Thus, such characteristics as education and training, actual labor market experience, and tenure may potentially be impacted by feedback effects. Measured labor market discrimination is likely to be underestimated to the extent that factors representing other dimensions of discrimination are controlled for.

The difficulties posed by these biases may be illustrated by a consideration of whether it is appropriate to include controls for occupation in analyses of discrimination. Their use tends to add considerably to the explanatory power of regressions (see below), though exactly how much depends on the precise form and amount of detail of the categories employed. If it is believed that individuals freely choose occupations according to their talents, tastes, and plans for labor force participation, differences in rewards for different types of work are appropriately included in the "explained" part of the earnings gap. If, on the other hand, occupational differences are caused by labor market discrimination and the adverse feedback effects it generates, occupational variables should be excluded. In the not implausible case that both discrimination and choice play a role, neither procedure will be unbiased. Indeed, what is called for is a two-equation model that explicitly treats occupation as endogenous.[3] Nonetheless, it is likely that specifications which exclude occupations (and other similar variables) yield an upper-bound estimate of discrimination and those which include such variables yield a lower bound.

With these issues in mind, we review the findings of the studies summarized in Table 1. It may be seen that, by the conventional measure, labor market discrimination has a substantial negative effect on the earnings of women and blacks.[4] Mean estimates of adjusted earnings ratios range from .62 to .81 for women and .67 to .81 for blacks, depending on specification. The mean unexplained portion of the gross differential ranges from 51 to 87 percent for women and from 39 to 66 percent for blacks. Table 1 also indicates the impact of alternative specifications. In the case of both gender and race differentials, the estimate of discrimination is considerably higher when a control for occupation is omitted.

The improved conceptualization of the role of labor market experience in explaining the gender differential (Jacob Mincer and

[2] Recently it has been suggested that the bias produced by imperfect measures of productivity could be eliminated by using a "reverse regression," in which the wage is the independent variable and an index of qualifications is the dependent variable. However, it appears that this method is not clearly superior and may be as much subject to bias as the traditional approach (Arthur Goldberger, 1984).

[3] Alternatively, a reduced form may be estimated in which occupational variables are excluded, but the underlying exogenous variables which determine them (for example, socioeconomic background) are included (Blinder).

[4] See Cain for a review of the findings for other minorities.

TABLE 1—SUMMARY OF CROSS-SECTIONAL STUDIES OF GENDER AND RACE DIFFERENTIALS

Control for			Means		
Occupation	Actual Experience	N^b	U_r	A_r^c	U_x
I. Gender (Women/Men)[a]					
No	No	8	.56	.62	.87
			(.13)	(.12)	(.10)
No	Yes	2	.65	.81	.56
			(.01)	(.01)	(.01)
Yes	No	6	.57	.70	.66
			(.14)	(.14)	(.10)
Yes	Yes	6	.59	.79	.51
			(.17)	(.12)	(.15)
II. Race (Black Men/White Men)					
	Hours of Work[d]				
No	No	4	.49	.67	.66
			(.04)	(.07)	(.16)
No	Yes	6	.60	.81	.47
			(.12)	(.11)	(.18)
Yes	Yes	1	.49	.80	.39

Source: Compiled with some additional calculations and small corrections from the studies summarized in Cain. Cain's table is in turn based on that presented in Donald Treiman and Heidi Hartmann (1981).

Note: Standard deviations are shown in parentheses.

[a] Includes results for all (black and white) or white workers. In the studies which present separate results for blacks and whites, U_r, A_r, and U_x tend to be higher for blacks than whites.

[b] The included studies were published between 1964 and 1981. Studies may be counted more than once if results are presented for different specifications or years. Citations of the studies are available upon request from the authors.

[c] Generally obtained by applying the findings for U_x based on regressions in which log earnings is the dependent variable to the observed earnings ratio (U_r).

[d] A study is included in this category if there is an explicit control for hours of work (annual, weekly, full-time/part-time) or if hours are implicitly controlled for by employing the wage rate as the dependent variable.

Solomon Polachek, 1974), as well as the more recent availability of appropriate information, mark a significant advance in our understanding of the causes of earnings differences by sex. As may be seen in Table 1, inclusion of actual (as opposed to potential) labor market experience considerably lowers the estimate of discrimination. It is important to note, however, that feedback effects have not been adequately taken into account by these studies (Blau, 1984). This suggests that, at least with regard to this consideration, labor market discrimination has been underestimated.

In the case of race differentials, it may reasonably be argued that discrimination has lowered the hours worked by black relative to white men (Cain). Blacks face substantially higher rates of unemployment and involuntary part-time employment. Their annual hours may also have been reduced by labor market discrimination in that it decreases the opportunity cost of temporary labor force withdrawals. Table 1 illustrates the substantial positive impact on the estimate of discrimination of omitting a control for hours worked.

In light of such sensitivity of the magnitude of discrimination to plausible alternative specifications of the underlying regression equations, and the fact that neither productivity nor discrimination itself is directly observable, some skepticism has arisen about the adequacy of this technique for measuring discrimination. Concern focuses on the possibility noted above that unmeasured nondiscriminatory factors could account for some or all of these differences. In the case of womeň, the traditional division of labor in the family may lower their motivation; commitment, and effort. For blacks, the poorer education they received, even when years of schooling are held constant, might be responsible in part for their lesser success. It is, however, likely that some unmeasured factors could work in the opposite direction.[5] Moreover, it is impressive that substantial unexplained race and sex pay differentials remain even in the most conservative estimates.

The conclusion from the cross-sectional studies that labor market discrimination plays an important role has been challenged recently by researchers using time-series data.

[5] For example, there is some evidence that, all else equal, the labor force group is more selective of those with higher wage offers among women (relative to men) and among black (relative to white) males (Blau and Andrea Beller, 1986). The issue of selectivity bias has been of particular concern in time-series studies (for example, Charles Brown, 1982; James Smith and Michael Ward, 1984).

Their studies show that the earnings gap between blacks and whites (James Smith and Finis Welch, 1986), as well as between women and men (June O'Neill, 1985; Smith and Michael Ward, 1984), has been narrowing, while relative human capital endowments of the relevant groups were changing correspondingly. Though the argument that the two developments were related is entirely convincing, it is not necessarily the case that the increase in human capital accounted for *all* of the observed improvement. There have been studies which suggest that equal opportunity legislation played a part (for example, Andrea Beller, 1979; Charles Brown, 1982). Further, a perceived decline in labor market discrimination may have helped to induce the very increase in human capital upon which these studies focus.

II. Other Sources of Evidence

Additional evidence of the importance of labor market discrimination is embodied in the significant number of court cases where individual employers have been found guilty of employment discrimination or have reached out-of-court settlements with the plaintiffs. Although these are at best "case studies," the number of such instances strongly suggests that they are more than isolated incidents. While the statistical techniques used to support legal charges of discrimination are quite similar to those reviewed above (demonstrating the wide impact of this approach), the argument of important omitted variables is less compelling in this context where it should be possible for firms to document the nondiscriminatory criteria which resulted in the observed race (sex) difference, if indeed such criteria exist.

Finally, it may be instructive to consider the direct evidence of discriminatory attitudes that has been accumulated by other social scientists. This research strikes us as particularly interesting in what is suggests about the nature of such attitudes. For example, a number of studies have found that both female and male college students gave identical papers higher ratings on such dimensions as value, persuasiveness, profundity, writing style and competence when respondents believed the author to be male rather than female. Similar findings have been obtained in studies requiring both women and men to evaluate the resumés of applicants for employment (for a review, see Virginia O'Leary and Ranald Hansen, 1982). Moreover, beliefs regarding sex differences in average ability and productivity-related behavior appear to be quite prevalent among managers (Benson Rosen and T. H. Jerdee, 1978).

Of course, the existence of such discriminatory attitudes does not prove that people act on them. It is, however, reasonable to suppose that employers who perceive the performance of members of a particular group as inferior would treat them accordingly without realizing that this might be costly for them. Some evidence in this regard is Ferber and Michele Teiman's (1980) finding that the acceptance rate for papers authored by women relative to those authored by men was higher in economics journals with double-blind refereeing than in those which had not adopted this practice.

III. Conclusions

A review of empirical studies suggests that even when fairly refined measures of productivity-related characteristics are held constant, blacks and women earn less than whites and men. We are inclined to accept this as evidence of labor market discrimination, although it is not possible at this point to determine the precise magnitude of the discriminatory wage effect. Future research could shed greater light on the process of discrimination if more effort were devoted to empirically testing various models of discrimination, and if it were recognized that a number of causes of discrimination may interact and reinforce each other. Our understanding of discrimination would also be enhanced if more were known about the mechanisms by which these outcomes are produced. Considerably more work is needed to understand fully the causes of race/sex differences in occupations, including the precise role played by discrimination in recruitment, hiring, job assignment, and promotion. The puzzling persistence of large racial differences in unemployment rates is also much in need of further research. Finally, estimation of feedback effects is an extremely important area which has been neglected in previous work.

In both conventional studies testing for discrimination and in those focusing on the other issues outlined above, we believe that researchers need to continue to explore all possibilities for getting more and better data on qualifications, performance and personnel practices, including information on individual enterprises where better measures of some of these variables may be obtained. Such information might be complemented by data on perceptions and attitudes from direct surveys of employers and workers as well as from experiments. Even with improved methods and better data it may be no easier to make progress towards consensus in the future than it has been in the past. In view of the pressing policy relevance of this research, however, it is as important as ever to continue trying.

REFERENCES

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957; 2nd ed., 1971.

Beller, Andrea H., "The Impact of Equal Employment Opportunity Laws on the Male/Female Earnings Differential," in C. B. Lloyd et. al., eds., *Women in the Labor Market*, New York: Columbia University Press, 1979, 304–30.

Blau, Francine D., "Discrimination Against Women: Theory and Evidence," in W. A. Darity, Jr., ed., *Labor Economics: Modern Views*, Boston: Kluwer-Nijhoff, 1984, 53–89.

______ **and Beller, Andrea H.,** "Trends in Earnings Differentials by Sex: 1971–1981," paper presented at the American Statistical Association Meetings, Chicago, 1986.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–55.

Brown, Charles, "The Federal Attack on Labor Market Discrimination: The Mouse That Roared?," *Research in Labor Economics*, Vol. 5, 1982, 33–68.

Cain, Glen G., "The Economic Analysis of Labor Market Discrimination: A Survey," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Amsterdam: North-Holland Press, forthcoming 1987.

Darity, William A., Jr., "The Human Capital Approach to Black-White Earnings Inequality: Some Unsettled Questions," *Journal of Human Resources*, Winter 1982, *17*, 72–93.

Ferber, Marianne A. and Teiman, Michele, "Are Women at a Disadvantage in Publishing Journal Articles?," *Eastern Economic Journal*, August-October 1980, *6*, 189–94.

Goldberger, Arthur, "Reverse Regression and Salary Discrimination," *Journal of Human Resources*, Summer 1984, *19*, 293–318.

Madden, Janice, "The Persistence of Pay Differentials: The Economics of Sex Discrimination," *Women and Work*, Vol. 1, 1985, 76–114.

Mincer, Jacob and Polachek, Solomon W., "Family Investments in Human Capital: Earnings of Women," *Journal of Political Economy*, March/April 1974, *82*, S76–S108.

O'Leary, Virginia E. and Hansen, Ranald D., "Trying Hurts Women, Helps Men: The Meaning of Effort," in H. Bernardin, ed., *Women in the Work Force*, New York: Praeger, 1982, 100–23.

O'Neill, June, "The Trend in the Male-Female Wage Gap in the United States," *Journal of Labor Economics*, January 1985, *3*, S91–S116.

Rosen, Benson and T. H. Jerdee, "Perceived Sex Differences in Managerially Relevant Behavior," *Sex Roles*, December 1978, *4*, 837–43.

Smith, James P. and Ward, Michael P., *Women's Wages and Work in the Twentieth Century*, Santa Monica: Rand Corporation, 1984.

______ **and Welch, Finis R.,** *Closing the Gap, Forty Years of Economic Progress for Blacks*, Santa Monica: Rand Corporation, 1986.

Treiman, Donald J. and Hartmann, Heidi I., *Women, Work, and Wages: Equal Pay for Jobs of Equal Value*, Washington: National Academy Press, 1981.

The Economics of Discrimination: Economists Enter the Courtroom

By ORLEY ASHENFELTER AND RONALD OAXACA*

Gary Becker's Ph.D. dissertation, subsequently published in 1957 as *The Economics of Discrimination*, is all the more remarkable when considered against the political and cultural background of the 1950's. Becker's book was written and published in a period when discrimination against black and women workers was legal in most states. Becker referred extensively to the scholarly literature on discrimination, but, with a couple of exceptions, his was a topic generally reserved for sociologists. Many things changed in the 1960's and the Equal Pay Act of 1963 and the Civil Rights Act of 1964 were both passed by the U.S. Congress. The economics of discrimination became a fashionable topic and, as so often happens in our discipline, several theorists attempted to lay claim to the ground that Becker had first staked out.

Becker also initiated the empirical study of the observable economic effects of discrimination: wage and income differences between black and white workers. The published scholarly studies of differences in wage rates between black, female, and other workers became increasingly sophisticated as better and more detailed data became available throughout the 1960's and 1970's. A particularly influential early paper for its clear methodological statement, use of microeconomic data, and detailed study of pay differences due to sex was Oaxaca's, published in 1973. It seems fair to us to observe that most of these early empirical studies were a response to the combination of new data available and a generally awakened interest in the scholarly study of the economics of discrimination. In particular, this work was not sponsored by government or private parties with a clear stake in its outcome.

Although *The Economics of Discrimination* has left a large scholarly legacy, we believe the empirical methods associated with the study of race and sex discrimination have had a still larger impact on practical matters. Our purpose in this paper is to give some small insight into how this early scholarly literature has ended up as a major factor in the litigation of many civil disputes where race and sex discrimination are alleged. We believe the simple concepts set out in *The Economics of Discrimination*, coupled with straightforward econometric tools in everyday use, have had a far larger impact on how Title VII (banning employment discrimination) of the 1964 Civil Rights Act is enforced through private litigation than is commonly understood by economists. Of course, whether the legal enforcement of bans on employment discrimination has had any measurable effect on the welfare of black or female workers compared to others is a topic of continued controversy, and we do not attempt to evaluate that issue here. Instead, our purpose is to show by example how economists reasoning about discrimination, and how it might be measured, have come to play a considerable role in the courtroom.

I. Economists Define and Measure Discrimination

The most basic idea in *The Economics of Discrimination* is that *market discrimination is defined* by a comparison of the wage rates of two groups, *W* and *N*, (a) as they are actually observed, and (b) as they would be observed in the absence of discrimination. For example, if *W* and *N* are perfect substitutes in production, in the absence of discrimination *W* and *N* would have the same wage rates. In this case, the difference be-

*Industrial Relations Section, Firestone Library, Princeton University, Princeton, NJ 08544, and University of Arizona, Tucson, AZ 87621, respectively.

tween the wage rates of the two groups W and N is a measure of discrimination.

More generally, if the observed wages of groups W and N are π_w and π_n, and if they would be π_w^0 and π_n^0 in the absence of discrimination, then the *proportionate market discrimination* against group N is

$$(1)\quad D=\left[\pi_w/\pi_n-\pi_w^0/\pi_n^0\right]/\left(\pi_w^0/\pi_n^0\right)$$
$$\approx \ln(\pi_w/\pi_n)-\ln\left(\pi_w^0/\pi_n^0\right).$$

D is the proportionate shortfall in the N to W wage ratio from what it would be in the absence of discrimination. It is clear that a similar definition of market discrimination is available for any market outcome, whether it be the number employed, hired, or discharged, or some other measure of compensation.

Since the wage ratio π_w/π_n is actually observed, implementing this definition of market discrimination is tantamount to specifying an empirical theory of wage determination that would be expected to prevail in the absence of discrimination. As Oaxaca observed, there are two natural alternative ways to do this. Suppose that it is agreed that some characteristics (in a vector) X determine pay in the absence of discrimination. Suppose further that for group W, the relationship between the wage, π_w, and these characteristics is of the form

$$(2)\qquad \ln\pi_w=\beta_w X+u,$$

where β_w is an unknown regression coefficient (vector) and u is a disturbance. Proportionate market discrimination is then approximately

$$(3)\quad D\approx\ln(\pi_w/\pi_n)-\beta_w(X_w-X_n),$$

where X_n and X_w represent the characteristics of the N's and W's being compared. In this setup, discrimination is measured as the difference between the observed proportionate wage difference between N's and W's and the proportionate wage difference that would be expected if N's were paid in the same way (i.e., according to the same regression function) as W's. If N's are the group against which discrimination is alleged, it is common, and perhaps most natural, to use this definition of market discrimination. In an analogous way, however, a measure of market discrimination that assumes W's are paid in the same way as N's in the absence of discrimination may also be constructed.

II. Economists in the Courtroom

Although these definitions of market discrimination have come to be almost commonsense to an economist, to lawyers they are irrelevant and hopelessly complex. For the legal mind, discrimination is above all an action taken by someone to the disadvantage of someone else *because of* their race or sex. Classic examples of such actions are the maintenance of separate racial lines of progression in a paper products factory, the refusal to hire women for certain blue-collar positions, or the maintenance of separate pay scales for black and white workers in similar jobs. These are examples of *disparate treatment*.

It is not hard to see that the appearance of disparate treatment is easy for an employer to eliminate without making any change in behavior at all. Differential hiring or pay scales may be supported by simply asserting that all hiring and pay is determined by merit, and merit is determined by employee supervisors. Who is to say whether employee supervisors *really* use "merit" in making their decisions? This obvious difficulty has lead the courts to recognize that some actions may be discriminatory because they have a *disparate impact* on the employment or compensation of one or more protected race/sex groups.

To most economists the insistence on finding "smoking gun" evidence of discriminatory actions, intent, or motivation seems quite irrelevant to determining whether market discrimination exists. In crude terms, for economists evidence of discrimination merely requires the presence of "unexplained" differences in compensation or employment. In practice, the economists' view has made considerable headway in the courts. This development has proceeded through a slow accretion of decisions that have placed more and more reliance on econometric methods in the determination of whether there is evidence of discrimination. The methods now presented to the courts look remarkably similar to the kind of studies that once appeared in the journals.

It is easy to see how a lawyer would object to the economists' procedures. First, note that *even if* it is agreed what factors (X) determine pay, equation (2) indicates that there are some unobserved factors (u) too. Even assuming the u's are uncorrelated with the X's and distributed identically for N's and W's, there is always a chance that the estimate of D ($\hat{D} = \ln(\pi_w/\pi_n) - \hat{\beta}_w(X_w - X_n)$) will be substantial because of sampling error alone. So long as the equation (2) does not fit perfectly, how can an economist say with certainty that $\hat{D}$ is not entirely a result of the chance configuration of the u's?

The answer to this question is, of course, that no such guarantee can be made. $\hat{D}$ has some sampling distribution and, as with ordinary econometric studies, only probability statements about the presence or absence of discrimination are possible.

In 1976 the Supreme Court started the move toward the use of the concept of statistical evidence in its decision in *Castaneda v. Partida*. The *Castaneda v. Partida* case involved the question of whether the grand jury selections in Hidalgo County, Texas, systematically underrepresented Hispanic Americans. The important feature of this decision is that no convincing anecdotal evidence demonstrating the intent to discriminate was presented. Yet the Court concluded that since grand jury selection was not by random device (which is used in the federal courts), the purely statistical evidence alone was sufficient to require rebuttal by the state.

It is very interesting to read how the Court reached the conclusion that the statistical disparity was enough "proof." In a long footnote, the Court presents a calculation that indicates the difference between the actual number of Hispanic grand jurors serving in Hidalgo County in an eleven-year period and the "expected" number of Hispanic grand jurors that would have served if there had been no discrimination. The expected number of Hispanic grand jurors is calculated by assuming that in the absence of discrimination, Hispanic jurors would be drawn in proportion to their representation in the population. This formulation is, of course, precisely that of an economist using equation (1) to define discrimination and using as equation (2) a model with a binomially distributed error term (u). Despite what some commentators have tried to suggest, the Court did not adopt a level of statistical significance for determining what constitutes proof. Instead, the Court stated:

> As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist.
>
> [p. 512, fn. 17]

The Court then proceeded to show that the difference between the actual and expected number of Hispanic grand jurors in this case was significantly different from zero at the 1 in 10^{140} level. The Court concluded that "the proof in this case was enough..." (p. 512).

The key part of the *Casteneda v. Partida* case is the explicit recognition that the presence of an error term in equation (2) does not automatically rule out the use of the econometric approach. Instead, using the econometric approach recognizes that, as always, little or nothing in life is certain, and risks must be taken. In a series of subsequent decisions, this basic idea has been elaborated to the point where evidence from stochastic models is now a regular feature in many different parts of the law, but especially in the analysis of discrimination.

A second objection to the economist's procedure is more subtle. Since there are omitted factors determining pay according to equation (2), what guarantees can be made that these factors are uncorrelated with a worker's race or sex? Might not omitted variable bias produce a finding that $\hat{D}$ is significantly different from zero when a properly specified version of equation (2) would not?

The answer to this question is, of course, that no economist can make the appropriate guarantee that there is no specification error in (2). Instead, the analysis will be more or less convincing according to (a) how much is well documented in general about models of the determination of pay, and (b) how well the particular study is documented and how complete it is. In other words, an economist will be more or less convinced by the findings of a particular nonexperimental study according to how well it is done!

Much to the surprise of some, the Supreme Court has very recently expressed much the same view. In the case of *Bazemore et al. v. William Friday et al.*, decided in July 1986, black workers in the North Carolina Extension Service alleged discrimination in pay that had begun before the Civil Rights Act of 1964 was extended to cover government workers (in 1972) and continued thereafter. The two lower courts had refused to accept regression studies offered by the black Extension Service workers to establish evidence of discrimination on the grounds that an appropriate regression analysis should include *all* measurable variables. The Supreme Court found this decision to be in error and provided the following more general guidance:

> ...it is clear that a regression analysis that includes less than "all measurable variables" may serve to prove a plaintiff's case... . Whether, in fact, such a regression analysis does carry the plaintiffs' ultimate burden will depend in a given case on the factual context of each case in light of all the evidence presented by both the plaintiff and the defendant... . [p. 331]

It would be difficult to provide a more concrete description of the way economists set about drawing their own conclusions in the give and take of the typical seminar.

A final legal objection to the use of the economists' measure of discrimination is that it need not reflect "actions" taken during the period when discrimination is "illegal." For example, suppose that $\hat{D}$ is measured when discrimination is "legal" in period $t-1$ as $\hat{D}_{t-1}$. Suppose also that $\hat{D}$ is measured again in period t, when it is illegal, as $\hat{D}_t$. Suppose further that $\hat{D}_{t-1}=\hat{D}_t$. To many lawyers the finding that $\hat{D}_{t-1}=\hat{D}_t$ means that any discriminatory "acts" that lead to $\hat{D}\neq 0$ must have occurred during the time when discrimination was legal. Therefore, it is argued, the finding $\hat{D}_t\neq 0$ does not imply "current" discrimination.

To economists this discussion seems completely to miss the point. Since $D\neq 0$ is defined by equation (1) to be the presence of discrimination, for an economist the finding $\hat{D}_{t-1}=\hat{D}_t\neq 0$ implies there was discrimination when it was legal *and* that this discrimination continues during the period when it is illegal!

Until the *Bazemore et al.* decision of July 1986, the legal community was completely divided on how this issue should be resolved. In the *Bazemore et al.* case, the lower courts found that before 1965, the North Carolina "Extension Service maintained two separate, racially segregated branches and paid black employees less than white employees." The Court of Appeals further acknowledged that the Extension Service had not eliminated all pay disparities in subsequent, actionable years, but claimed that it was not the employers duty under the law to do so. The Supreme Court's decision stated:

> The error of the Court of Appeals with respect to salary disparities created prior to 1972 and perpetuated thereafter is too obvious to warrant extended discussion: that the Extension Service discriminated with respect to salaries *prior* to the time it was covered by Title VII does not excuse perpetuating that discrimination after the Extension Service became covered by Title VII. To hold otherwise would have the effect of exempting from liability those employers who were historically the greatest offenders of the rights of blacks. [p. 328]

This decision is, of course, equivalent to applying the economists definition of market discrimination to the factual findings of the *Bazemore et al.* case.

There are many further applications of the economists' methods that could also be incorporated into the settlement of disputes over allegations of discrimination. For example, once a class of plaintiffs has prevailed in a discrimination suit and established a legitimate claim of discrimination, it is necessary to establish the compensatory damages owed to this "class." A common approach is to set aside a sum *$K*, and then to hold numerous minitrials to determine how these funds should be allocated to the individual members of the class. The legal costs associated with this procedure are usually substantial.

An obvious and far less expensive alternative procedure is suggested by the economist's approach. Assuming that the presence of discrimination has not altered the compensation that W's received (an assumption that Title VII of the 1964 Civil Rights Act maintains by requiring that no worker's pay be reduced to eliminate the presence of discrimination) implies that $(1+\hat{D})\pi_n$ multiplied times the number of N workers is a natural measure of the "back pay" owed to N's. There remains the issue of how these funds are to be allocated among the individual class members. A natural procedure is to assign payments by predicting from (2) the salary that each class member would otherwise have received, and then to assign the difference between each class member's predicted and actual salary as the compensatory award. There are problems with this approach, however, because it will assign some individuals negative awards when their actual pay is higher than their predicted pay. The simple procedure of assigning awards only where they are positive is no remedy to this defect, because such a procedure will produce an estimate of total compensatory damages greater than the estimate of discrimination in (3). The precise implementation of these procedures is, therefore, still a matter where alternative schemes deserve exploration. In view of the cost of alternative procedures, however, we believe these issues will also eventually be resolved.

III. Conclusion

Although written some thirty years ago, the underlying framework in *The Economics of Discrimination*, coupled with some simple, modern econometric methods, has become the standard form by which the litigation of disputes over allegations of race and sex discrimination proceeds. It is easy for economists to understate the magnitude of this accomplishment. After all, the setup in *The Economics of Discrimination* seems to an economist approaching it today almost like commonsense. To the legal profession, however, the definition of market discrimination in Becker's book is far from natural. This is surely a testament to the power and simplicity of the ideas in *The Economics of Discrimination*.

REFERENCES

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, October 1973, *14*, 693–709.

Castaneda, Claudio v. Partida, Rodrigo, No. 75-1552, *U.S. Supreme Court Reports*, 51 L Ed 2d, pp. 498–526, 1976.

Bazemore, P. E. et al. v. Friday, William C. et al., No. 85-93, *U.S. Supreme Court Reports*, 92 L Ed 2d, pp. 315–43, 1986.

REAL BUSINESS CYCLE THEORY: WHAT DOES IT EXPLAIN?[†]

Allocative Disturbances and Specific Capital in Real Business Cycle Theories

By STEVEN J. DAVIS*

Macroeconomists have grown disenchanted, albeit in varying degree, with business cycle theories that either posit unexplained nominal wage and price rigidities or rely on misperceptions about nominal variables as a key driving force. One response to this disenchantment has been a concentration of research effort on real business cycle theories. Aside from dissatisfaction with competitor theories, the very visible oil price shock episodes of the 1970's lent plausibility to the view that exogenous "real" disturbances cause large fluctuations in aggregate economic activity.

This essay on real business cycle theory considers the role of allocative disturbances in aggregate economic fluctuations when human and physical capital are specialized. By "allocative disturbances," I mean events that impinge on the economy by inducing a costly, time-consuming reallocation of specialized resources. At least since the publication of Ricardo's *Principles* in 1817, economists have recognized some of the potentially important aggregate consequences of allocative disturbances. Ricardo writes:

> A great manufacturing country is peculiarly exposed to temporary reverses and contingencies, produced by the removal of capital from one employment to another.... [C]onsiderable distress, and no doubt some loss, will be experienced by those who are engaged in the manufacture of such [adversely affected] commodities; and it will be felt not only at the time of the change, but through the whole interval during which they are removing their capitals, and the labour which they can command, from one employment to another. [1951, p. 263]

As important sources of allocative disturbances, Ricardo suggests the capricious nature of tastes for nonagricultural commodities, the sector-specific effects of tax changes, the uneven sectoral effects of technological innovations, and, especially, the impact of war on patterns of demand and channels of trade. To this list, recent experience suggests we add exogenous disturbances to the supply of intermediate inputs. Despite these varied sources of allocative disturbances, neither specialization of capital and labor nor nontrivial reallocation processes figure prominently in currently popular business cycle theories, real or otherwise. This state of affairs reflects a belief that the effects of allocative disturbances on aggregate economic activity are of second-order importance and, also, the difficulty of incorporating nontrivial specialization and reallocation technologies into tractable general equilibrium models of economic fluctuations.

My main purpose in this essay is to stimulate further theoretical and empirical research directed towards understanding the consequences of allocative disturbances. I argue that the aggregate consequences of allocative disturbances play a large role in economic fluctuations and that the incorporation of nontrivial specialization and reallocation technologies into real business cycle models is a promising research strategy. I discuss some recent evidence on real-world business cycles that indicates the quantitative significance of fluctuations in labor real-

[†] *Discussant*: Robert G. King, University of Rochester.

[*] Graduate School of Business, University of Chicago, Chicago, IL 60637.

location, convey some important insights that emerge from analyses of multisector business cycle models with costly reallocation technologies, and discuss the implications of our experience with oil price shocks for real business cycle theories. I also show how some of the chief criticisms directed against real business cycle theories can be addressed by considering the implications of the costly reallocation of specialized capital and labor. To begin, I introduce some evidence on the nature of short-run unemployment rate fluctuations.

For many persons, especially job losers, labor mobility involves measured unemployment spells. This basic observation motivates search and matching models that interpret unemployment as a nonemployment activity distinct from leisure consumption. This nonemployment activity facilitates the reallocation of specialized labor towards a more productive pattern of employment. Christopher Flinn and James Heckman (1983) provide formal econometric evidence that the state of unemployment facilitates labor mobility. Given these considerations, it is natural to investigate unemployment rate data for evidence that fluctuations in the pace of labor reallocation are an important aspect of business cycles.

In forthcoming work, I find that fluctuations in the pace of labor reallocation across jobs are the largest component of short-run unemployment rate fluctuations. This conclusion emerges by combining three observations: (*i*) Short-run increases (decreases) in the unemployment rate coincide with increases (decreases) in the flow rate of persons into the unemployment pool *and* the flow rate of persons out of the unemployment pool. The unemployment flow rates are calculated from *Current Population Survey* (*CPS*) data on unemployment by duration. This pattern of comovement among the inflow, outflow, and unemployment rates holds for every recession in the postwar U.S. economy. (*ii*) The contemporaneous correlation between changes in the labor force participation rate and the flow rate of persons out of the unemployment pool is virtually zero. This observation suggests that the pattern described in (*i*) does not reflect cyclic movements of persons into and out of the labor force. (*iii*) Increases in permanent separations account for the largest fraction of short-run unemployment rate increases during recessions; in particular, they account for a larger fraction than increases in temporary separations. Permanent and temporary separations are calculated from *CPS* data, available since 1967, on unemployment by reason. The raw numbers indicate that increases in permanent and temporary separations contribute roughly equal amounts to unemployment rate increases during recessions. However, several pieces of evidence summarized in my forthcoming paper indicate that the raw numbers overstate the importance of temporary layoff unemployment. Many economists claim that the 1970's witnessed unusually large changes in the composition of labor demand, so that broad conclusions about the relative importance of permanent and temporary separations drawn from post-1967 experience are open to question. Pending the accumulation of further evidence, a more conservative interpretation of the data might be appropriate—namely, that increases in both permanent and temporary separations contribute significantly to short-run increases in unemployment. In a nutshell, recessions are characterized by an unusually rapid rate of job loss by employed persons *and* by an unusually rapid rate of attachment of unemployed persons to new jobs.

What are the implications of this finding for real business cycle theories? First, to the extent that firm-worker matches entail relationship-specific capital, this finding constitutes direct evidence that the reallocation of labor resources across meaningfully defined sectors fluctuates greatly over the business cycle. The sectoral shifts literature contains ample evidence that the reallocation of labor resources across broad industrial sectors also fluctuates greatly over the business cycle. The phenomenon of large fluctuations in the pace of costly labor reallocation lies outside the framework of real business cycle models à la Finn Kydland and Edward Prescott (1982) or John Long and Charles Plosser (1983). In defense of their models, one might argue that whether people substitute into leisure or into mobility during recessions is far less interesting than the observation that they substitute out of market work. Furthermore, the abstraction of lumping all nonmarket activities into a single category, dubbed leisure, is an extremely powerful simplifying device. The counterargument is simply that the cost of dispensing with this abstraction will be recouped many times over by a greater understanding of economic fluctuations. The payoffs come in four varieties: (*i*) a richer perspective on why people substitute into or out of market work; (*ii*) an additional class of observable disturbances that drive economic fluctuations; (*iii*) an understanding of how allocative disturbances and what we traditionally call aggregate disturbances interact in interesting, nonobvious ways; and (*iv*) the development of additional economic mechanisms that propagate real disturbances over time and across sectors. I address each point below.

Second, the finding that reallocative unemployment increases during recessions is not at variance with the better-known finding that temporary layoff unemployment, as a fraction of total unemployment, increases during recessions. Temporary layoff unemployment fluctuates greatly over the business cycle, but it constitutes a small fraction of total unemployment even during recessions. Over the years 1967 to 1985, temporary layoff unemployment averages 14.7 percent of total unemployment, reaching a low of 10.7 percent in 1973 and a high of 21.2 percent in 1975. (These figures, based on *CPS* data, include persons who report themselves as laid off with no definite recall date. Caveats about the overstatement of temporary layoff unemployment apply here.) Sharp cyclic movements in temporary layoff unemployment partly account for the pattern of comovement among the inflow, outflow and unemployment rates. Sharp cyclic variation in temporary layoff unemployment is compatible with real business cycle models that abstract from specialization and reallocation. See, for example, Gary Hansen's (1985) analysis of the interaction between labor supply nonconvexities and incentives to substitute leisure intertemporally.

In a multisector setting with labor supply nonconvexities and barriers to short-run labor mobility, temporary layoff unemployment can arise from disturbances that have little or no effect on the aggregate (i.e., cross-sectional average) marginal product of labor. Similarly, adverse aggregate disturbances with uneven effects across sectors can produce more temporary layoff unemployment than equal magnitude disturbances with even effects across sectors. To see these points, consider a disturbance that increases the gap between labor's marginal product in different sectors and, perhaps, lowers the average marginal product. Suppose that the average marginal product remains above the value of leisure, but the marginal product falls below the value of leisure (for some workers) in the most adversely impacted sectors. If the disturbance is transitory, or some probability is attached to this possibility, persons in the most adversely impacted sectors may rationally choose to consume leisure rather than switch sectors. Large sector-switching costs imply the optimality of waiting for improvement in sectoral marginal product rather than working or moving. James Hamilton's (1986) analysis illustrates equilibrium "waiting-type" unemployment in a fully articulated multisector model. The only frictions in his model are a time cost of changing sectors and a restriction that persons supply zero or one unit of labor. The import of these remarks on temporary layoff unemployment in a multisector economy becomes apparent by recalling that sectoral experiences during recessions are sharply dissimilar. Some sectors exhibit high cyclical sensitivity, some do not. This dissimilarity, coupled with labor supply nonconvexities and a modest willingness to substitute leisure intertemporally, explains why adverse aggregate disturbances with small effects on labor's average marginal product cause big substitutions out of work and into temporary layoff unemployment (i.e., leisure).

Third, the finding that fluctuations in the pace of labor reallocation constitute a large fraction of unemployment rate fluctuations highlights the importance of the central controversy in the sectoral shifts literature: Are fluctuations in the pace of labor reallocation, and the concomitant fluctuations in the unemployment rate, driven by the arrival pattern of allocative disturbances over time? Or, are these fluctuations driven largely by the timing of what we traditionally call aggregate disturbances? I identify the sectoral shifts hypothesis with an affirmative answer to the first question and various reallocation timing hypotheses with an affirmative answer to the second question. David Lilien (1982), Katharine Abraham and Laurence Katz (1986), my papers (1986a, b, forthcoming), and other contributors to the sectoral shifts literature seek to address these and related questions. In my view, a full resolution of the issues surrounding these questions requires confrontation between the data and fully articulated multisector models incorporating nontrivial reallocation technologies.

The sectoral shifts hypothesis has attracted attention precisely because it departs sharply from traditional notions about the driving forces behind aggregate economic fluctuations—Ricardo's observations notwithstanding. The benefits of intersectoral labor mobility depend on the disparity in conditions across sectors—in particular, on cross-sectoral differences between current plus expected returns to labor. Allocative disturbances that create sufficiently large sectoral disparities induce a reallocation of labor towards more productive employment. In the short run, allocative disturbances decrease the economy's productive potential because of the time cost of labor mobility, the loss of specific capital, retraining costs, and the difficulty of reestablishing an efficient pattern of matches between firms and workers and among workers. Thus, a class of disturbances that implies little or no effect in models that abstract from specialization implies important aggregate effects in an economy with specialized capital and labor. Fischer Black (1982), myself (1986b), and Hamilton (1986) model this phenomenon.

Hamilton's analysis shows how seemingly small allocative disturbances cause large fluctuations in aggregate output and unemployment when labor is specialized. Hamilton assumes that energy, a nonproduced good, is a complement to one of the produced goods in utility functions or a complement to labor in one of the sectoral production functions. Individuals supply labor to only one sector at a time and sector switching entails one period out of work. In this setting, an exogenous decrease in the supply of energy directly triggers a reduction in the output of goods that use energy in consumption or production. If the energy supply decrease is perceived as persistent, then output declines even further in the short run than implied by the complementarity assumption, because workers suffer unemployment while moving out of the energy-using sector. Here, the short-run output decline is bounded not by the value of energy's share in total output, but by the output share of goods that use energy in production or consumption. This result implies that real business cycle models that consider the allocative effects of oil price shocks can explain the large output and employment declines that followed postwar oil price increases.

Our experience with oil price shocks provides the most persuasive evidence that real disturbances cause large aggregate fluctuations. Real business cycle models that abstract from specialization and reallocation seem incapable of explaining the apparent magnitude of the response to oil price shocks, especially the relatively mild shocks that occurred during the pre-OPEC period. Hamilton (1983) and many others document the magnitude of the responses, while Hamilton (1983, 1985) documents the exogenous character of the oil price shocks. My paper (1986a) and Prakash Loungani (1986) give evidence supporting interpretations that stress the allocative effects of oil price shocks. When the relative price of oil is held fixed, Loungani shows that Lilien-type dispersion measures have no residual explanatory power for unemployment rate fluctuations. His finding holds for the postwar period and the 1900–29 period. My paper showed that oil price shocks explain much of the time-series variation in the pace of labor reallocation (as proxied by a Lilien-type dispersion measure) and do so in a way predicted by the sectoral shifts hypothesis. The empirical findings by Hamilton, Loungani, and myself carry a cogent message: allocative disturbances and specific capital will be critical features of successful real business cycle theories.

Beyond offering a coherent explanation for the effects of oil price shocks, the sectoral shifts view points to important economic propagation mechanisms. A careful development of economic mechanisms that propagate disturbances over time and across sectors marks much of the best real business cycle research; Kydland and Prescott's development of time-to-build technologies for durable goods and Long and Plosser's treatment of capitalistic production possibilities stand as clear examples. Propagation mechanisms stressed by the sectoral shifts view include specific human capital investment, the time costs of prospecting for a job or a better match, and other barriers to short-run labor mobility. These propagation mechanisms imply a testable prediction that serves to gauge their importance in aggregate economic fluctuations: the contemporaneous unemployment response to an allocative disturbance depends, in part, on past patterns of labor reallocation. An allocative disturbance that reinforces past patterns of labor reallocation exacerbates skill, location, and informational mismatches between workers and firms. In contrast, an allocative disturbance that reverses past patterns of labor reallocation mitigates skill, location, and informational mismatches. An unfavorable allocative disturbance, in the sense implied above, increases unemployment sharply relative to a favorable allocative. A favorable allocative disturbance may even decrease unemployment.

In my forthcoming paper I test this prediction using annual data from 1924 to 1985 by constructing indexes for the current direction of labor reallocation relative to past directions. The indexes are cross-sectoral covariance measures:

$$(1)\quad \sigma_{t,j}^{H}=\sum_{i=1}^{N}\left(\frac{x_{it}}{X_t}\right)(\Delta_1 x_{it}-\Delta_1 X_t)\times(\Delta_j x_{i,t-1}-\Delta_j X_{t-1}),\quad j=1,2,\dots,J,$$

where x_{it} = employment in sector i at time t, X_t = aggregate employment at time t, $\Delta_j x_{it} = \ln x_{it} - \ln x_{i,t-j}$, and N = the number of labor-market sectors broken down by industrial classification. $\sigma_{t,j}^{H}$ indexes the time t direction of labor reallocation over a one-period horizon relative to the $t-1$ direction over a j period horizon. Relatively high (low) values of $\sigma_{t,1}^{H}, \sigma_{t,2}^{H},\dots$ indicate that the time t direction of labor reallocation reinforces (reverses) past patterns of labor reallocation. The sectoral shifts hypothesis predicts a positive partial correlation between directional indexes and the economywide unemployment rate.

In unemployment regressions that include Lilien-type dispersion measures, unanticipated monetary disturbances, government purchases, and proxies for the financial intermediation collapse during the 1930's as explanatory variables, I find strong support for this prediction. Coefficients on the directional indexes exhibit the predicted sign and are highly statistically significant. Furthermore, the range of unemployment rate fluctuations accounted for by the directional indexes exceeds the range accounted for by the Lilien-type dispersion measure and its lags. The econometric results survive several specification changes and are reinforced by more casual inspection of the data. The United States exhibited tremendous sectoral reallocation in the demobilization following World War II, but the directional indexes show that the pattern of labor reallocation simply reversed the pattern during the preceeding war years; unemployment was very low. In contrast, 1932 was also a year of tremendous sectoral labor reallocation, but the directional indexes show that the pattern sharply reinforced the pattern during the preceeding years; unemployment was very high. These empirical findings are important for three reasons: they constitute independent support for the sectoral shifts view, they highlight the significance of the propagation mechanisms stressed by the sectoral shifts view, and they suggest that allocative disturbances other than energy price shocks have caused large aggregate fluctuations.

The evidence on oil price shocks and directional indexes provides compelling support for the sectoral shifts view, but a strong form of the sectoral shifts hypothesis leaves much, perhaps most, of the short-run fluctuations in the pace of labor reallocation and the concomitant movements in unemployment unexplained. Reallocation timing hypotheses point to other reasons for fluctuations in the pace of labor reallocation. Richard Rogerson (1986), myself (1986b), and Robert Topel and Laurence Weiss (1985) model three different reallocation timing mechanisms.

Rogerson shows how the timing of secular movements between sectors with differential trend productivity growth rates depends on the relative cyclical sensitivities of the two sectors. In his model, workers discount future returns and costs, and they face a fixed cost of sector switching that varies across workers. Naturally, low-cost workers move first, but each worker delays his costly move until the discounted lifetime returns to moving equal the discounted lifetime returns to a further delay plus the moving cost. If the shrinking sector is more (less) cyclically sensitive, then the intersectoral flow of workers accelerates (decelerates) when the cross-sectoral average marginal product is below (above) trend. In this setting, none of the variation in the pace of labor reallocation stems from the arrival pattern over time of allocative disturbances.

My paper (1986b) considers a multisector model with allocative disturbances and cross-sectoral average marginal product disturbances that arrive randomly over time. Sector switching takes a period of time, so that the opportunity cost of labor mobility varies with the value of foregone production. Convex labor force adjustment costs incurred by firms imply that the sectoral labor force distribution converges slowly in the direction of the target distribution. In this setting, average product disturbances do not change the target distribution, but they alter the timing of reallocation towards the target by affecting the opportunity cost of labor mobility. Favorable aggregate disturbances decrease the pace of labor reallocation and unemployment, while unfavorable aggregate disturbances increase the pace of labor reallocation and unemployment. Here, the reallocation timing mechanism is the intertemporal substitution of labor mobility. The strength of this intertemporal substitution effect depends on parameters describing the mobility and adjustment cost technologies and on the degree of diminishing returns to sectoral inputs. The intertemporal substitution effect operates whenever labor reallocation involves foregone production and shocks affecting average product occur.

Topel-Weiss identify a different intertemporal substitution mechanism that influences the timing of labor mobility. They consider a multisector model with sector-specific human capital investment that is costly, irreversible, and risky. The riskiness stems from the possibility that future events will alter the pattern of productive opportunities and reduce the *ex post* return to sector-specific investment. Recognizing this possibility, agents may choose to delay sector-specific investment and mobility, because they anticipate a resolution of uncertainty about the pattern of productive opportunities. In this setting, specific human capital investment and mobility become more or less attractive as uncertainty about future productive opportunities shrinks or grows.

The analyses by Rogerson, myself and Topel-Weiss suggest the rich potential for interesting interaction between allocative and aggregate disturbances in the determination of economic fluctuations. They show how explicit modeling of specialization and reallocation technologies yields new mechanisms by which aggregate disturbances influence the timing of labor reallocation and cause economic fluctuations.

To close, I make one final remark. This essay emphasizes how consideration of specialization and reallocation technologies can improve our understanding of the driving forces behind economic fluctuations and the mechanisms that propagate disturbances over time and across sectors. The ideas expressed in this paper have yet to be integrated into a real business cycle model as ambitious in scope as the models of Kydland-Prescott and Long-Plosser. This effort deserves a high priority on the business cycle research agenda.

REFERENCES

Abraham, Katharine and Katz, Lawrence F., "Cyclical Unemployment: Sectoral Shifts or Aggregate Disturbances?," *Journal of Political Economy*, June 1986, *94*, 507–22.

Black, Fischer, "General Equilibrium and Business Cycles," NBER Working Paper No. 920, 1982.

Davis, Steve J., (1986a) "Allocative Disturbances and Temporal Asymmetry in Labor Market Fluctuations," University of Chicago Working Paper Series No. 86–38, 1986.

______, (1986b) "Sectoral Shifts and the Dynamic Behavior of Unemployment: A Theoretical Analysis," University of Chicago Working Paper Series No. 86–35, 1986.

______, "Fluctuations in the Pace of Labor Reallocation" in *Carnegie-Rochester Conference Series on Public Policy*, forthcoming 1987.

Flinn, Christopher J. and Heckman, James J., "Are Unemployment and Out of the Labor Force Behaviorally Distinct Labor Force States?," *Journal of Labor Economics*, January 1983, *1*, 28–42.

Hamilton, James, "Oil and the Macroeconomy since World War II," *Journal of Political Economy*, April 1983, *91*, 228–48.

______, "Historical Causes of Postwar Oil Shocks and Recessions," *Energy Journal*, January 1985, *6*, 97–116.

______, "A Neoclassical Model of Unemployment and the Business Cycle," University of Virginia, 1986.

Hansen, Gary, "Indivisible Labor and the Business Cycle," *Journal of Monetary Economics*, November 1985, *16*, 309–27.

Kydland, Finn and Prescott, Edward, "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Lilien, David, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *90*, 777–93.

Long, John and Plosser, Charles, "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

Loungani, Prakash, "Oil Price Shocks and the Dispersion Hypothesis, 1900–1980," University of Rochester Center for Economic Research Working Paper No. 133, 1986.

Ricardo, David, *On the Principles of Political Economy and Taxation*, Cambridge: Cambridge University Press, 1951.

Rogerson, Richard, "Sectoral Shifts and Cyclical Fluctuations," University of Rochester, 1986.

Topel, Robert and Weiss, Laurence, "Sectoral Uncertainty and Unemployment," University of California-Los Angeles, 1985.

Sectoral vs. Aggregate Shocks In The Business Cycle

By JOHN B. LONG, JR. AND CHARLES I. PLOSSER*

Comovement among a wide variety of economic activities is an essential empirical characteristic of business cycles. The received tradition in macroeconomics is to interpret this comovement as evidence of a common aggregate disturbance. The observation that the nominal price level and commodity outputs moved together, at least pre-World War II, further suggested to many economists that fluctuations in aggregate demand is the relevant common disturbance that is the key to understanding business cycles.

Recently a class of stochastic neoclassical growth models has been adapted to analyze business cycles. These models focus on the self-interested responses of economic agents to productivity disturbances and are commonly referred to as real business cycle models.[1] In multisector versions of these models (1983), we showed that even if random productivity shocks are independent across sectors, agent's choices will cause comovement of activity measures from different sectors. Thus, observed comovements do not logically dictate the presence of a common or aggregate disturbance.

The purpose of this paper is to look directly at the comovement in commodity outputs in an attempt to determine the extent to which it can be characterized as resulting from a common aggregate shock or from a more diverse set of independent disturbances.

I. Background

In most business cycle theories with random shocks, output growths for period t can be interpreted as functions of (*i*) shocks that occur in period t, and (*ii*) decisions based on information (prices, resource stocks, past shocks, etc.) available prior to period t. To make precise the distinction between period t shocks and prior influences, the period t shocks are defined as the time t *innovations* in the actual variables (for example, aggregate demand, total factor productivities by industry) that characterize the shocks. Conditional expectations of these variables are part of the prior influences on period t output growths. Thus, it is the innovations—the unexpected part—of output growths in period t that most directly reflect the period t shocks.

A focus on output innovations is especially important in distinguishing the influence of many cross-sectionally independent shocks from that of one or two aggregate shocks that affect all outputs. As we (1983) show, the *unconditional* joint distribution of output growths in a many-shock model can be very similar to the *unconditional* distribution from an aggregate shock model.

If the major source of randomness in outputs is in fact one or two aggregate shock processes, then the correlation matrix of output innovations will have large off-diagonal elements and will be consistent with one or two unobservable common factors obtained from a factor analysis. If, on the other hand, the major source of randomness is many independent shocks, then the correlation matrix of innovations will not be consistent with a small number of factors. If the diverse shocks are industry specific (instead of, for example, regionally specific), then the off-diagonal elements of the correlation matrix will be approximately zero.

To empirically identify monthly output innovations, we use the following specification

$$\mathbf{y}_t = \mathbf{S}d_t + A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + A_3\mathbf{y}_{t-12} + \mathbf{e}_t, \tag{1}$$

*William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627. Plosser's research is supported by the National Science Foundation.

[1] See our 1983 paper, Robert King and Plosser (1984), Finn Kydland and Edward Prescott (1982), and King, Plosser, and S. Rebelo (1986).

where **y** is a vector of monthly output growth rates and **e** is a vector of disturbances. **S** is a matrix of seasonal (monthly) means and d_t is a dummy variable that has a one in the jth row when t is the jth month. This specification is consistent with a version of the model in our 1983 paper and when fitted to our sample yields serially uncorrelated residuals.

II. Empirical Results

Our data consist of the continuously compounded rates of growth for thirteen commodity output groups from the industry structure decomposition of the Index of Industrial Production (IIP). The data are not seasonally adjusted. The monthly data are summarized in Tables 1 and 2. In Table 1 the amount of comovement among these industries is summarized by the average pairwise correlation of each industry with all other industries. For example, the Food industry has an average correlation of .41 with the other industries. These numbers show a fair amount of cross-sectional correlation (i.e., comovement), even at the monthly level.

Since the data are highly seasonal, it is informative to remove the most obvious seasonal fluctuations. The second column of Table 1 shows that after removing monthly means, outputs display substantially less comovement. Thus a large portion of the cross-sectional correlation arises from similar seasonal patterns in different industries. Also, the R^2 from the monthly dummy regressions (R^2S), reported in the second column of Table 2, indicate the substantial explanatory power of the monthly means. The R^2s range from .24 to .84 with the median being .67.

TABLE 1—AVERAGE PAIRWISE CORRELATIONS MONTHLY, 2/48–12/81

			Innovations[c]		
Industry	Unadj.[a]	Adj.[b]	*RMS*	*RMSE1*	*RMSE2*
Food	.41	.07	.17	.04	.04
Leather	.56	.20	.26	.10	.02
Textiles	.51	.12	.18	.09	.03
Fuels	.26	.06	.09	.06	.06
Chemical	.55	.23	.26	.08	.04
Plastics	.55	.22	.21	.04	.05
Paper	.59	.18	.26	.06	.04
Wood	.52	.11	.18	.06	.04
Metals	.42	.15	.14	.07	.06
Machines	.53	.28	.28	.05	.06
Home Durables	.50	.18	.23	.06	.05
Glass	.39	.20	.20	.07	.04
Vehicles	.16	.12	.15	.06	.07

[a] For each industry, average pairwise correlation with other industries from the cross-sectional correlation matrix of unadjusted monthly growth rates.

[b] From the correlation matrix of seasonally adjusted monthly growth rates.

[c] From the correlation matrix of VAR residuals and from 1 and 2-factor decompositions of the matrix. *RMS* for each industry is the root mean square of correlations with other industries. *RMSEn* for each industry is the root mean square error in the n-factor representation of correlations with other industries.

After removing monthly means, most of the output growth rates remain slightly serially correlated. There may also be some lagged cross correlation. In order to isolate innovations, we estimate system (1) using the thirteen industry groupings. This yields a set of serially uncorrelated innovations. The R^2V in Table 2 is the R^2 associated with each equation in system (1). The root mean square of the off-diagonal elements of the correlation matrix of the innovations for each industry is in Table 1. Root mean square is simply another way of measuring the extent of comovement. It weights large correlation (both positive and negative) more than averaging.[2]

In order to measure the potential contribution of an aggregate disturbance, we do a simple factor analysis on the innovations. Factor analysis is a statistical procedure that decomposes a set of random variables into unobserved common factors and a set of unique disturbances. Thus, a factor model attributes all of the comovement to the common factors, which we interpret as aggregate disturbances. This assignment maximizes the estimated contribution of aggregate disturbances, since the true model may be driven by disaggregated shocks that happen to be correlated. It is also important to keep in mind that the estimated common factors are not constrained to be any observable aggre-

[2] The average pairwise correlation of the innovations is very similar to those of the deviations from monthly means.

TABLE 2—MONTHLY GROWTH RATES, 2/48–12/81

				Factor Models[d]	
Industry	σ^a	R^2S^b	R^2V^c	R^21F	R^22F
Food	3.6	.81	.88	.12	.12
Leather	7.6	.84	.91	.30	.68
Textiles	8.0	.84	.90	.11	.27
Fuels	2.8	.24	.41	.01	.03
Chemical	2.8	.54	.78	.34	.47
Plastics	5.6	.62	.73	.21	.19
Paper	3.8	.82	.91	.34	.37
Wood	6.6	.69	.78	.14	.17
Metals	6.8	.37	.52	.07	.10
Machines	3.1	.62	.75	.40	.36
Home Durables	4.7	.67	.82	.27	.28
Glass	4.0	.72	.81	.16	.26
Vehicles	14.1	.30	.53	.09	.07

[a] Standard deviation of monthly growth rate of output in percent per month.
[b] R^2 from regression of monthly growth rate on 12 seasonal dummy variables.
[c] R^2 from monthly growth rate VAR with seasonal dummy variables.
[d] 1 and 2-factor models of residuals from the VAR. The R^2s are fractions of VAR residual variance explained by the common factors.

gate disturbance and thus may overstate the potential explanatory power of any popular observable shock. We are therefore measuring an upper bound of the explanatory power of aggregate disturbances.

In Table 2, the R^2, or explanatory power of the common factors, of both a one- and two-factor model of the innovations is reported. The range is from .01 for Fuels to .40 for Machines with median of .16 for the one-factor model. By expanding to two factors, the median increases to .26. This second factor, however, is primarily explaining Leather, Textiles and to a lesser extent Chemical and Glass. Thus, this factor does not seem well characterized as a common disturbance.[3] In Table 1 the effect of the common factor can be ascertained from the root mean square correlations associated with the unique disturbances. These measure the extent to which the off-diagonal elements of the innovation correlation matrix deviate from zero after taking into account the estimated common factor. These numbers are quite small compared to those for the innovations and adding a second factor does not yield very much improvement.[4]

Finally, we investigate the extent to which our estimated common shock is capable of explaining some measure of and the innovation to aggregate industrial output. There are several ways one might proceed. We take the raw sectoral innovations and weight them by the industry's share in the overall Index of Industrial Production. We then compute the explanatory power of the common factor for this measure of the innovation in the aggregate IIP. For the one-factor model, the common factor accounts for approximately 47 percent of the variance of this aggregate innovation.[5]

These results are consistent with the existence of an aggregate disturbance, but one with limited explanatory power for sectoral industrial outputs. This is most obviously reflected in the limited amount of comovement in the monthly innovations. On the other hand, if one wants to explain something closer to an aggregate, our evidence suggests that the aggregate shock model is a glass that is either half empty or half full depending on your point of view.

It might be argued that looking at monthly innovations reduces the potential role for common shocks if these shocks influence some sectors only with some delay. The only way to adequately address this is to put more structure on the problem than we have so far. Time aggregating to quarterly data is sometimes suggested as a way of mitigating this phenomenon. On the other hand, moving to quarterly observations may result in mislabelling some portion of sectoral shocks that have propagated to other sectors within the quarter as common disturbances. Our view is that if one is interested in isolating shocks to sectoral productivities, the ap-

[3] In fact, as more factors are added, each tends to increase the R^2 of only one or two industries at a time.

[4] Although the second factor is statistically significant.

[5] This strategy is not equivalent to as computing the innovations to the IIP from the lags of IIP, but is analogous to computing the innovation from lags of each of the sectors. We leave it to the reader to determine if this measure is of any interest or not.

propriate time interval is the one that most closely corresponds to the production interval. On these grounds, it would seem that the monthly time interval is the better choice.

Nevertheless, we have conducted a parallel investigation to the monthly results presented above using quarterly data. The results are largely what one would expect from time aggregation of serially correlated variables. Standard deviations are larger and seasonal dummies explain a high proportion of the variance. After removing monthly means, however, the average pairwise correlations remain substantial. They range from .08 for Food to .48 for Chemical with a median value of .37. Again, this is to be expected if most of the comovement comes from propagation of industry-specific shocks or from a common shock that impulses different sectors at different points in time.

The pairwise correlations of the quarterly innovations are very close to those implied by aggregating the estimated monthly equations. This suggests that the monthly and quarterly results are quite compatible with one another. Using quarterly data, the one-factor model explains from 10 percent for Food to 77 percent for Chemical of the quarterly innovation variance with the median value being 41 percent. As in the monthly case, a one-factor model seems to capture the bulk of the comovement among the industries since the *RMSE*s for the factor models is generally quite small, although formal tests would indicate that more factors were required.

III. Summary

The data we have investigated suggests that the explanatory power of a common aggregate disturbance for industrial outputs is significant, but not very large for most industries. This result arises even though our factor analysis procedure attributes all correlations among industry innovations to a common factor. If any part of the observed comovement of industry output innovations is attributed to independent disaggregate influences like regionally specific shocks, then the implied explanatory power of an aggregate factor is less than we have estimated.

REFERENCES

King, R. G. and Plosser, C. I., "Money Credit and Prices in a Real Business Cycle," *American Economic Review*, June 1984, *74*, 363–80.

______, ______, **and Rebelo, S.,** "Production, Growth and Business Cycles," manuscript, 1986.

Kydland, F. and Prescott, E. C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Long, J. B. and Plosser, C. I., "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

Is Consumption Insufficiently Sensitive to Innovations in Income?

By LAWRENCE J. CHRISTIANO*

A basic fact about U.S. macroeconomic data is that consumption is a much smoother time series than income. A classic explanation for this is a simple version of Friedman's permanent income hypothesis (SPIH). According to the SPIH, an innovation to current income causes households to revise their consumption plan by the annuity value of that innovation.[1] The annuity value is computed under the assumption of a constant interest rate.[2] When U.S. income data are modeled as the sum of a linear trend and a covariance stationary process, then innovations to income do not affect the long-run income outlook and their annuity value is smaller than the innovation itself (Angus Deaton, 1986, equations (7)–(8)). Thus, with this model of income the SPIH predicts, correctly, that consumption is smoother than income in the sense that consumption's innovations are a fraction of income's.

However, a growing number of researchers are attracted to the view that U.S. income data can be represented as a positively autocorrelated process in first differences. In this view, an innovation to income produces a change in the long-run outlook for income and has an annuity value greater than the innovation itself. Then, the SPIH has the strongly counterfactual implication that consumption is less smooth than income. Put differently, measured U.S. consumption is insufficiently sensitive to innovations in income, relative to the SPIH and a first-difference specification for income. This is an implication emphasized by Deaton.

The analysis below suggests that the reason for the SPIH's counterfactual implication is its fixed interest rate assumption. Using a parametric version of the standard model of economic growth. I show that very small movements in interest rates are enough to induce an empirically plausible amount of comsumption smoothing.

I study a version of Gary Hansen's 1985 model. (Another model that formalizes the argument in this paper is in my paper with Martin Eichenbaum and David Marshall, 1987.) My model is formulated so that equilibrium income is a positively autocorrelated process in first differences. In this model, disturbances to income result from permanent marginal productivity shocks. Consequently, a positive innovation to income signals not only that households' long-run ability to consume has risen, but that the return to investment (i.e., the interest rate) has also risen. The increase in the long-run ability to consume, by itself, motivates households to substantially increase consumption today. This income effect on consumption is partially offset by the substitution effect as households take advantage of the increased return to saving. The combined income and substitution effects have the consequence that the smoothness of consumption relative to income implied by my model is about what is observed. This is noteworthy because the model's parameter values are chosen to match averages of U.S. time-series data, with

[*] Research Department, Federal Reserve Bank of Minneapolis, Minneapolis, MN 55480. I have benefited from conversations with S. Rao Aiyagari, David Backus, V. V. Chari, Martin Eichenbaum, Gary Hansen, Martha Hellander, Patrick Kehoe, and Edward Prescott. The financial support of the National Science Foundation is gratefully acknowledged. The views expressed here are mine alone.

[1] Lars Peter Hansen (1985) and Thomas Sargent (1986) study a general equilibrium, representative agent growth model which rationalizes the SPIH. The essential feature of their model is that the technology shock only affects the average product of capital; the marginal product is constant. Also, utility is quadratic in consumption, and hours are not in the model.

[2] If $\theta(L)y_t = C(L)\varepsilon_t$ is the ARMA representation for income y_t and r is the real (assumed fixed) rate of interest, then the annuity value of an innovation ε_t in y_t is $rC[(1+r)^{-1}]/\theta[(1+r)^{-1}]$ (see Angus Deaton).

second-moment properties playing a minimal role. A serious evaluation of the model's explanation for consumption smoothing cannot ignore its implication for other aspects of the data. I therefore also explore some of these.

I. The Model

At date $t=0$, a representative agent chooses decision rules for c_t, h_t, and dk_t to maximize[3]

$$(1)\quad E_0 \sum_{t=0}^{\infty} \beta^t \{\ln c_t - \gamma h_t\}, \quad \text{for} \quad \gamma > 0,$$

subject to the technology

$$(2)\quad c_t + k_t - [(1-\delta)/n] k_{t-1} = n^{-\theta}(z_t h_t)^{(1-\theta)} k_{t-1}^{\theta}.$$

Here c_t denotes consumption, h_t hours worked, k_t end-of-quarter stock of capital, and dk_t capital investment in quarter t. The expression on the right side of (2) is output y_t. This is assumed to be related to k_{t-1}, h_t, and a technology shock z_t by a Cobb-Douglas production function. The variables k_t and dk_t are related by $k_t - [(1-\delta)/n]k_{t-1} \equiv dk_t$. All variables are measured per capita. Assumed constant are the parameters n, the gross growth rate of the population, and δ, the rate of depreciation of a unit of capital.

The growth rate x_t of the technology shock is assumed to be covariance stationary with a first-order autoregressive structure. In particular,

$$(3)\quad z_t = z_{t-1}\exp(x_t), \quad x_t = \mu + \rho x_{t-1} + \varepsilon_t, \quad \text{for} \quad |\rho| < 1,$$

where ε_t is white noise. According to (3), the average growth rate of the technology shock is $\mu/(1-\rho)$, with first-order autocorrelation ρ. In the model, c_t, y_t, k_t, and dk_t grow at the same rate as z_t. On average, per capita hours h_t do not grow, which is roughly in accord with postwar U.S. experience.

To derive the model's implications for the stochastic properties of its endogenous variables, the decision rules are needed. Because obtaining these exactly is complicated, instead I obtain approximations. (For details, see my 1986b paper.)

The analysis also requires the equilibrium rate r_t at which a unit of consumption can be transformed risklessly from t to $t+1$. This is defined as

$$(4)\quad 1 + r_t = [\partial u(c_t, h_t)/\partial C_t] / [\beta E_t \partial u(c_{t+1}, h_{t+1})/\partial C_{t+1}].$$

Here $u(c_t, h_t) = \log(c_t) - \gamma h_t$ and $C_t \equiv N_t c_t$, where N_t is the population in quarter t. When σ_ε is small, the average value of r_t is $[\mu/(1-\rho)](n/\beta) - 1$. This is roughly the sum of the economywide rate of consumption growth $[n\mu/(1-\rho) - 1]$ and the subjective rate of time discount $(\beta^{-1} - 1)$.

II. Parameter Values

To deduce the model's quantitative implications, values must be assigned to its parameters. I choose these:[4] $\rho = -.077$, $\mu = .0035$, $\gamma = .0026$, $n = 1.00324$, $\beta = .99$, $\delta =$

[3] One interpretation of the immortal representative agent is in S. Rao Aiyagari (1986). In a framework that nests mine, he shows how a utility function expressed in terms of per capita consumption and hours, like (1), can summarize preferences in an economy that has a growing number of overlapping generations of people with finite lives and operative bequest motives. Like mine, Aiyagari's is a model with uncertainty in which per capita consumption, capital, and output grow on average.

[4] The approximate decision rules for k_t and h_t implied by these values are $k_t = z_t\exp\{9.75 + .9494[\log(k_{t-1}) - \log(z_{t-1}) - 9.75] - .9441(x_t - .00325)\}$ and $h_t = \exp\{5.78 - .4540[\log(k_{t-1}) - \log(z_{t-1}) - 9.75] + .5201(x_t - .00325)\}$. (See my 1986 paper for details.) Decision rules for the model's other variables are derived using (2) and the definition of the production function.

.018, $\theta = .39$, and $\sigma_\varepsilon = .019$. The value of n is the average quarterly growth in the quality-adjusted, working-age population in 1952–84. With this value, $\delta = .018$ is required if the gross investment series implied by dk_t is to resemble the gross investment series published by the U.S. Department of Commerce. The value of β is from Finn Kydland and Edward Prescott (1982). Values for θ, γ, and $\mu/(1-\rho)$ are chosen to roughly match the model's implications for the average values of h_t, c_t/y_t, and k_t/y_t with their empirical counterparts in U.S. data for 1956:II–1984:I. The implied averages (and empirical values) for these variables are 323.9 (320.4), .72 (.72), and 11.32 (10.58), respectively. The values of ρ, μ, and σ_ε are based on an analysis of the time-series properties of z_t, which can be measured using data on y_t, k_t, and h_t given the values assigned to θ and n.

Consumption is defined as public and private consumption of goods, services, and the services of the stock of durables. The stock of capital is defined as the stock of public and private equipment and structures plus the stock of consumer durables plus public and private residential capital. Capital investment is defined to conform to the definition of the capital stock. I use G. Hansen's (1984) time-series on hours worked measured in efficiency units. Variables are converted to per capita terms by the working-age population, measured in efficiency units. The risk-free rate is proxied by the *ex post* real return on three-month U.S. Treasury bills. (For further details on the data and this methodology for choosing parameter values, see my 1986a paper.)

III. Relative Smoothness of Consumption

Here I describe the dynamic properties of the model from two perspectives. First, the model's shock response function is used to deduce the model's implication for the relative smoothness of consumption and income. Then several of the model's unconditional second-moment properties are examined. These provide an alternative, complementary, measure of smoothness.

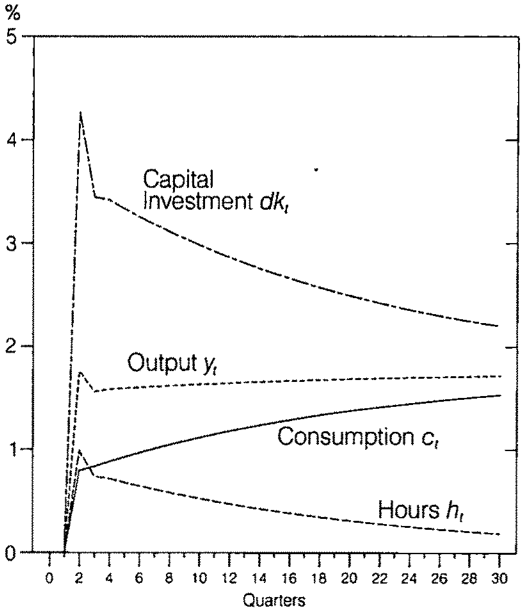


FIGURE 1. RESPONSE OF MODEL TO $\varepsilon_2 = .019$: PERCENT DEVIATIONS FROM STEADY-STATE BASELINE

Figure 1 shows the first 30 quarters' responses of c_t, h_t, dk_t, and y_t to a one standard deviation innovation in the growth rate of the technology shock z_t in period 2 given that the system is on a steady-state growth path in $t = 0,1$ (i.e, $\varepsilon_2 = .019$, $\varepsilon_t = 0$ for $t = 0,1,3,4,\dots$). The curves are the quarterly percentage deviations in these variables from a baseline scenario in which $\varepsilon_t = 0$ for $t = 0,1,2,3,\dots$.

With the assumed stochastic structure of z_t, an innovation to z_t is 92.85 percent $[100/(1-\rho)]$ permanent. Thus, in period 2, z_t jumps 1.9 percent above its baseline growth path, then declines to a path 1.76 percent above the baseline. After the shock, all the model's variables except r_t and h_t end up 1.76 percent above the baseline. As Figure 1 shows, consumption rises only gradually to this higher growth path. In particular, households choose not to adjust consumption immediately, as the SPIH—which only recognizes an income effect—implies. This reflects households' desire to delay consumption when the return to investment is high (the substitution effect). Thus, capital investment responds strongly.

On Figure 1, note the early spikes in the responses of dk_t, h_t, and y_t. This reflects the fact that 7.15 percent of the initial 1.9 percent jump in z_t is only temporary. The lack of a spike in c_t reflects the small response of consumption to a temporary disturbance, which explains the pronounced spike in capital investment. Hours also respond fairly strongly to the temporary component in the productivity shock (as in G. Hansen, 1985).

The ratio of the jumps in consumption and income in period 2 is .32; that is, consumption's innovation is about 32 percent of income's. The empirically measured value of this ratio is 33 percent.[5]

The shock response of r_t is not on Figure 1 because it is so small. In the steady state, $r_t = 1.01667$. After the shock, r_t rises to 1.01728, then declines monotonically back to 1.01667. Thus, the effect on the interest rate is a negligible six one-hundredths of a basis point. Evidently, this model generates an empirically plausible degree of smoothness in consumption with only very little variation in the interest rate.

Next, I report smoothness properties of the model based on unconditional second moments. I refer to these measures of smoothness as *volatility*. Table 1 reports measures of the volatility of c_t, dk_t, h_t, and r_t relative to that of y_t as well as the volatility of y_t itself. The volatility of c_t, dk_t, and y_t is the standard deviation of the log of the first difference of these variables. The volatility of h_t and r_t is the standard deviation of their levels. The relative volatility measures are the ratios of these to the volatility of y_t. All standard deviations are computed for variables predicted by the model to be covariance stationary. Means and standard deviations for the volatility measures are computed by simulating 1,000 sets of 112 observations from the model. In each simulation, the decision rules of the model are solved with initial conditions on a steady-state growth path and ε_t's drawn independently from a normal random number generator with mean zero and standard error .019.

TABLE 1–SELECTED SECOND-MOMENT PROPERTIES

	Model Simulations[b]		
Statistic[a]	Mean	(Standard Deviation)	U.S. Estimates[c]
σ_c/σ_y	.44	(.031)	.49
σ_{dk}/σ_y	2.56	(.170)	1.91
σ_h/σ_y	378.23	(93.80)	669.59
σ_r/σ_y	.0777	(.0196)	2.24
σ_y	.0176	(.0012)	.0115
Er_t	.017	(.0009)	.0024
$\rho_{r,\Delta c}(0)$	.533	(.031)	.085
$\rho_{\Delta c,\Delta c}(1)$	.059	(.108)	.271
$\rho_{\Delta y,\Delta y}(1)$	−.119	(.093)	.361

[a] σ, E are the volatility and mean, respectively, of the indicated variable; $\rho_{u,v}(\tau)$ is the correlation between $u(t)$ and $v(t-\tau)$, $\tau = 0,1$; and $\Delta u(t)$ denotes $\log u(t) - \log u(t-1)$.

[b] 1,000, each 112 quarters long.

[c] 1956:II–1984:I.

Table 1 also shows empirical estimates of volatility measures for the U.S. economy. Note that consumption's relative volatility is .49. This is quite close to the model's prediction, which is only about 1.6 standard deviations lower. Thus—relative to this model, but in striking contrast to the SPIH—if there is a puzzle it is that the empirical relative variability of consumption is too high, not too low.

IV. Other Implications of the Model

The model does less well on other dimensions. Table 1 shows that the empirical measures for both investment and hours, for example, are more than three standard deviations from the model's predictions.

The most substantial evidence in Table 1 of a mismatch between the model's implications and the data is that for the risk-free return. The empirical measure of its relative variability is 110.32 standard deviations

[5] I estimate this by dividing the standard deviation of the innovation to consumption by the corresponding quantity for income, as implied by a three-lag vector autoregression in $c_t - c_{t-1}$ and $y_t - y_{t-1}$. My estimate of 33 percent is consistent with Deaton's estimate of 50 percent since my measure of y_t includes capital income, whereas his only includes labor income. Labor income is a fairly steady 66 percent of GNP. (See my 1986a paper, fn. 2.1.)

higher than the model's prediction. Also, the empirical correlation between the risk-free rate and consumption growth is 14.45 standard deviations below that correlation in the model, and the empirical average of the risk-free rate is 16.07 standard deviations below the model's average. It is not clear whether these discrepancies reflect shortcomings of the model or of my empirical measure of the risk-free rate.[6]

Another thing the model explains less well is the serial correlation in Δc_t and Δy_t. [$\Delta u_t \equiv \log(u_t) - \log(u_{t-1})$.] For example, as expected given the small movements in r_t, Δc_t is virtually uncorrelated with lagged Δc_t, Δy_t, Δdk_t, r_t, and h_t. The corresponding empirical quantities are all larger. (Only the correlation with lagged Δc_t is in Table 1.) What is perhaps more surprising is that the serial correlation properties of y_t closely match those of z_t with capital accumulation seemingly playing a small role. Thus, the model's first-order autocorrelation of Δy_t is $-.119$, or 5.34 standard deviations below the empirical value. Not surprisingly, part of the reason this model can match the observed relative smoothness of consumption is this negative serial correlation. For example, with $\rho = .2$ but all other parameters, including $\mu/(1-\rho)$, unchanged, the first-order serial correlation of Δy_t averages the empirically plausible .349, with standard deviation .084. Here, however, a consumption innovation is 51 percent of an income innovation and the volatility of consumption is 66 percent that of income, with standard deviation .038. Although these numbers are higher than the corresponding empirical values, they are considerably lower than what would be implied by the SPIH. As before, this is brought about by very small movements in the interest rate.

[6] For example, while T-bills may be close to risk free, the average household cannot borrow much, if at all, at this rate. In addition, the return on T-bills reflects not just their function of transferring consumption intertemporally, but also their function of providing liquidity. The model abstracts from the latter.

REFERENCES

Aiyagari, S. Rao, "Overlapping Generations and Infinitely Lived Agents," Research Department Working Paper 328, Federal Reserve Bank of Minneapolis, 1986.

Christiano, Lawrence J., (1986a) "Why Does Inventory Investment Fluctuate So Much?," paper presented at Catholic University of Lisbon/University of Rochester Conference on Real Business Cycles, Lisbon, Portugal, 1986.

______, (1986b) "Dynamic Properties of Two Approximate Solutions to a Particular Growth Model," Federal Reserve Bank of Minneapolis, 1986.

______, **Eichenbaum, Martin and Marshall, David,** "The Permanent Income Hypothesis Revisited," Federal Reserve Bank of Minneapolis, 1987.

Deaton, Angus, "Life-Cycle Models of Consumption: Is the Evidence Consistent with the Theory?," NBER Working Paper 1910, 1986.

Hansen, Gary D., "Fluctuations in Total Hours Worked: A Study Using Efficiency Units," University of Minnesota Working Paper, 1984.

______, "Indivisible Labor and the Business Cycle," *Journal of Monetary Economics*, November 1985, *16*, 309–27.

Hansen, Lars Peter, "Econometric Modeling of Asset Pricing under Rational Expectations," paper presented to Fifth World Congress of the Econometric Society, 1985.

Kydland, Finn E. and Prescott, Edward C., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Sargent, Thomas J., "Equilibrium Investment under Uncertainty: Measurement Errors and the Investment Accelerator," University of Minnesota, 1986.

ARBITRATION AND THE NEGOTIATION PROCESS[†]

Arbitrator Behavior

By ORLEY ASHENFELTER*

It is rarely worth specifying the contractual obligations in an ongoing agreement so completely as to cover every contingency. This means that disputes may arise in any ongoing economic relationship. When the parties have made specific investments that cannot be recaptured, it will often be the case that the consequences of an unresolved dispute are costly for both parties. The arbitration of disputes by a third party is intended to settle disputes in a way that avoids these costs.

In a labor agreement the costs of unresolved disputes may be dramatic, as when there is a strike, or they may evolve more slowly as the steady erosion of morale and productivity in the workplace. Arbitration systems are often used to resolve labor disputes, perhaps because ongoing employment relationships are so likely to contain specific (human capital) investments. It would be incorrect to suggest that arbitration systems are used only for the resolution of labor disputes, however. Commercial contracts often involve ongoing relationships where specific investments have been made and these contracts also often contain provisions for the arbitration of disputes (see Paul Joskow, 1987). Moreover, the resolution of disputes that end up in the courtroom, whether commercial or otherwise, bears much in common with their resolution by arbitration.

In a civil suit and in an arbitration proceeding, the disputing parties present their cases to a third party for a binding decision. In both cases the parties may negotiate a settlement whose terms are influenced to some extent by what it is expected the neutral would otherwise decide. A key difference between the arbitration of disputes and their resolution in a court, however, is in the nature of the fact finder. An arbitrator is typically a professional who is selected at least in part by mutual agreement of the parties to a dispute. The evidence suggests that a key determinant of the parties preferences for an arbitrator is usually the extent of the arbitrator's "experience" in deciding related arbitration cases (see David Bloom and Christopher Cavanagh, 1986). In the courts, however, juries are selected *because* they have little or no experience with the nature of a particular dispute, and the jurors are certainly never expected to cumulate "experience" by deciding a number of related cases. This suggests that arbitration appeals to the parties because it resolves their disputes without exposure to the greater risk associated with a court decision.

At the same time, a key feature of the *ex ante* acceptability of jury decisions is their unpredictability. Indeed, it is the inability to predict with certainty the outcome of a jury trial that defines what is meant by "fairness." A major finding that is emerging from the research on arbitrator behavior is that arbitrator decisions are also statistically *exchangeable*; that is, arbitrator decisions contain an unpredictable component that may be characterized by a probability density function.[1] It seems likely that it is the ex-

[†] *Discussants*: Robert Gibbons, Massachusetts Institute of Technology; Charles Plott, California Institute of Technology.

*Director, Industrial Relations Section, Firestone Library, Princeton University, Princeton, NJ 08544.

[1] A set of random variables $y_1,\dots,y_n$ is said to be K-exchangeable if the joint distributions of any $K < n$ of these random variables are the same. So long as a panel of jurors (or arbitrators) is K-exchangeable, for example, then any deterministic rule for aggregating K

changeability of arbitrator decisions that also leads to the continued acceptability of arbitration systems.

The discovery that actual arbitrator decisions are statistically exchangeable has resulted primarily from a series of simple empirical analyses and considerable data collection. The first section of this paper provides detailed explanations of two simple empirical examples that I believe most readers will find convincing as evidence for the arbitrator exchangeability hypothesis. Section II provides an explanation for why arbitrator exchangeability provides so good a statistical model for arbitrator decisions. The basic idea for this model is simple: The parties to an arbitration decision are always allowed to express their preferences in the selection of the arbitrator who will handle their case. Each party will naturally rule out arbitrators whose historical decisions are unfavorable to their position. Arbitrators who have taken extreme positions relative to their colleagues are thus excluded from future selection by either one party or the other. Knowing this, the strategy of a successful (i.e., enduring) arbitrator is to provide decisions that are forecasts of the decisions *other* arbitrators will make in similar situations. This is the only systematic strategy that keeps an arbitrator's decisions from looking aberrant. Arbitrators who follow this strategy thus make decisions that have the appearance of forecast errors; indeed, they *are* forecast errors.

I. Evidence of Arbitrator Behavior

Table 1 contains the basic facts on the operating characteristics of an arbitration system used to resolve disputes over compensation among public safety officers throughout the state of New Jersey. Under *conventional arbitration*, the parties present their cases to an arbitrator who fashions whatever award seems most reasonable. Under *final-offer arbitration*, each party must present an offer for selection by the arbitrator without compromise. Under the New Jersey statute, the parties opt for conventional arbitration when they mutually agree to adopt this procedure, while final-offer arbitration is used if they do not mutually agree to something else. As in most states, more than two-thirds of the compensation disputes in New Jersey are resolved by the parties without arbitration. Final-offer arbitration is used four to five times as frequently as conventional arbitration.

TABLE 1—FINAL-OFFER ARBITRATION IN NEW JERSEY POLICE DISPUTES: UNION OFFERS, EMPLOYERS OFFERS, AND AWARDS IN CONVENTIONAL ARBITRATION CASES[a]

	(1)	(2)	(3)	(4)	(5)
1978	7.14	6.55	5.01	7.41	32
1979	8.29	8.59	6.51	8.51	35
1980	8.54	8.26	5.70	8.27	27

Source: Tabulation of arbitration reports, state of New Jersey, presented in my article with Bloom (1984).
Note: Col. 1: Mean Union Offer; Col. 2: Mean Conventional Award (in other disputes); Col. 3: Mean Employer Offer; Col. 4: Predicted Mean of Conventional Awards; Col. 5: Employer Wins (percent of cases).
[a]Union and employer offers and the fact finder recommendations are expressed as proposed percentage changes in compensation.

The data in Table 1 raise an immediate puzzle, and it is the resolution of this puzzle that has lead to evidence for the arbitrator exchangeability hypothesis. Unlike the expectations of many, employers have won only about one-third of the final-offer decisions in New Jersey. What accounts for this unbalanced win-loss record?

A clue to solve the puzzle is available through a comparison of the sample statistics on the offers made by the union and employer bargainers (cols. 1 and 3, Table 1) with the sample statistics on the conventional awards (col. 2). In each of the years for which data are presented, it is obvious that the mean union offer was closer to the

juror's (or arbitrator's) decisions will lead to a stable stochastic distribution of awards. An interesting analysis of rules for the aggregation of juror preferences that (implicitly) uses the assumption of K-exchangeability is Alvin Klevorick et al. (1984). I find it fascinating that the concept of exchangeability, which is so intertwined with the definition of statistical behavior, is also closely related to the appearance of "fair behavior."

mean conventional arbitration award than was the mean employer offer. Using the conventional arbitration awards as a benchmark, it is clear that the union offers were, on average, more "reasonable" than the employer offers. Early on, it was suggested that perhaps this accounts for the union success in win-loss ratios and that a simple model where arbitrator decision making is characterized as stochastic, but independent of the institutional setting (conventional vs. final-offer arbitration), may unify the puzzling findings that characterize Table 1. The model I shall describe is more than a simple response to the data in Table 1, however, because it generates additional testable cross-equation (econometric) restrictions that may well be rejected. It is thus a candidate for explanation of the known facts that may explain still other data.

To proceed, suppose that arbitrator wage increase decisions, w, may be characterized as a draw from some distribution with density function $f(w|x)$, and parameters x. The simplest scheme is then to suppose that under conventional arbitration the arbitrator simply mandates the preferred award w. This implies that the parameters of the distribution of arbitrator preferences may be estimated directly from the observations on the conventional arbitration awards. Under this hypothesis an uncontaminated estimate of the mean of arbitrator preferences for wage increases in 1980 is, from Table 1, simply the mean of conventional arbitration awards in that year; that is, 8.26 percent.

How are we to imagine that an arbitrator will make a decision under final-offer arbitration? The simplest scheme is for an arbitrator to first prepare a preferred award, given the facts (x) of the case. Next, the arbitrator may compare w against w^e, the employer proposal, and w^u, the union proposal, and select whichever of these proposals is closer to w. The basic idea is demonstrated in Figure 1, where the density $f(w|x)$ is plotted. The point midway between w^e and w^u is $\frac{1}{2}(w^e+w^u)$. Since an arbitrator is expected to select whichever parties proposal is closer to the arbitrator's preferred award, w^e is selected if an arbitrator is drawn whose preferred award is less than $\frac{1}{2}(w^e+w^u)$. Alternatively, w^u is selected if an arbitrator is drawn whose preferred award is greater than $\frac{1}{2}(w^e+w^u)$. It follows that the probability of selecting the employer offer (P_e) is simply $P_e = F(\frac{1}{2}(w^e + w^u)|x)$, which is the shaded area under the curve in Figure 1 to the left of the point $\frac{1}{2}(w^e + w^u)$ on the horizontal axis. Increases in w^e or w^u thus increase the probability that the employer's offer is selected.

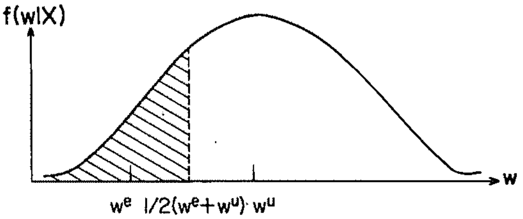


FIGURE 1

In order to approach the data in Table 1 it is necessary to be more concrete. Suppose, therefore, that the distribution of arbitrator preferences is normal with mean μ and standard deviation σ. Then the probability an employer's offer is selected is given by the probit function $P_e = F[\frac{1}{2}(w^e + w^u)/\sigma - \mu/\sigma]$, and μ and σ may be estimated from paired data on the mean of the offers and the identity of the winner. Estimates of μ and σ are thus available from the two separate, independent sets of data representing conventional and final-offer arbitration cases.

Column 2 of Table 1 contains the estimates of μ from the conventional arbitration cases while column 4 contains the estimates of μ from the probit function for final-offer arbitration cases. The "actual" and "predicted" means of arbitrator preferences from these two sets of data are remarkably close, and certainly not significantly different by conventional tests. (The estimated standard deviations, which are not reported, also match quite closely, and are around 2 percent.) These data clearly support the cross-equation restrictions implied by the arbitrator exchangeability hypothesis.

A second set of data covering eight years of final-offer arbitration cases in the Iowa public sector is described in Table 2. Final-offer arbitration is used in Iowa only when

TABLE 2—FINAL-OFFER ARBITRATION IN IOWA: EMPLOYERS OFFERS, FACT FINDER RECOMMENDATIONS, AND ACTUAL AND PREDICTED WIN-LOSS PERCENTAGES (Shown in Percent)

	(1)	(2)	(3)	(4)	(5)
All Years	7.54	5.96	4.89	65.5	61.1
1976[a]	10.61	6.18	5.67	100	80
1977	8.26	5.22	5.52	72.7	74
1978	13.89	5.08	5.57	0	76
1979	9.01	6.19	6.68	100	82
1980	10.89	9.44	8.95	66.7	56
1981	–	7.65	–	–	–
1982	6.91	3.64	5.14	75	57
1983	4.84	3.51	1.50	42.9	44

Source: Tabulation of arbitration reports, state of Iowa, presented in my paper with James Dow and Daniel Gallagher (1986).

Note: Cols. 1–3 are defined in Table 1 Note: Col. 4: Actual Employer Wins; Col. 5: Predicted Employer Wins.

[a] Union and employer offers and the fact finder recommendations are expressed as proposed percentage changes in compensation.

the parties opt to eliminate the provision that a fact finder (selected by the parties) is first asked to provide a recommended, but nonbinding compensation award. There are thus relatively few cases of final-offer arbitration in Iowa. The summary statistics for these cases also contain a puzzle, however. Again, win-loss ratios are unbalanced, but in Iowa employer offers have been accepted in two-thirds of the cases. How are we to account for this puzzle, *and* the reversal of arbitrator behavior from that found in New Jersey?

A key to the solution of this puzzle is found by comparing the mean of the union and employer offers with the mean recommended award in other arbitration cases in each year in Iowa. In each of the years, 1976–82, the employer offers were typically closer to the recommended awards elsewhere than were the union offers. By this standard, the employers were typically more "reasonable" in Iowa than were the unions. Consistent with the arbitrator exchangeability hypothesis, in five of the six years in which there were awards the employer offers were more likely to be accepted than the union offers. Remarkably, in 1983 the union offers were slightly closer to the recommended awards than the employer offers, and in that year the union offers were accepted more frequently also.

Column 5 of Table 2 provides the predicted values of P_e using μ and σ estimated from the independent data on fact finder recommendations. Although hardly exact, these predictions are impressive further confirmation of the arbitrator exchangeability hypothesis.

II. The Arbitrator Exchangeability Hypothesis

There is a simple rationale for the finding that arbitrators appear to be statistically exchangeable that is consistent with the anecdotal evidence of arbitrator behavior as well. At any given date we may imagine that the facts of a particular arbitration case are known to the parties and the arbitrator, and indicated by X_i. After all cases have been arbitrated there will be some population regression function

$$w_i = bX_i + \varepsilon_i$$

by which we may relate the arbitrator's decision w_i to the facts of the case X_i. If the weight b to be given to the facts were known, each arbitrator could simply elect to assign the decision $w_i = bX_i$, in which case $\varepsilon_i = 0$. Certainly no arbitrator could be accused of bias by following such a practice as they would all make identical decisions in the same circumstances. Arbitrators would certainly be exchangeable if this were to occur, but the distribution of their decisions would be degenerate. This kind of arbitrator behavior would no doubt generate negotiated settlements, but only because the settlement (bX_i) was effectively being dictated.

In fact, however, the population regression coefficient b is not observable at the time the arbitrator prepares an estimated b_j of b and makes the award

$$w_i = b_j X_i,$$

so that $\varepsilon_i = (b_j - b)X_i \neq 0$. In a single cross section, the variability in arbitrator awards will, therefore, look like forecast errors. Good arbitrators will be those for which $E(b_j - b)$

$= 0$ and $E[(b_j - b)\dot{X}_i] \neq 0$, where the expectation is taken over time. Indeed, arbitrators who do not satisfy these criteria will not be selected by the parties and must eventually leave the business. Thus, the variability of arbitrator awards among surviving arbitrators will have the same properties as well behaved forecast errors. The important point is that the arbitrator's preferred awards will resemble a stable stochastic process independent of the institutional setup in which the arbitrator participates.

It is important to emphasize that arbitrator exchangeability need not characterize all observed arbitrator behavior. Arbitrator exchangeability is the limiting behavior that would be observed if information collection were costless. In practice, however, the parties must have some incentive to continue to collect the information that allows them to determine whether an arbitrator might be partial toward or against their own case. It is the potential gain that comes from selecting an arbitrator who provides a party a slight probabilistic edge that gives the incentives for information collection that should drive arbitrator behavior toward exchangeability. The arbitrator exchangeability hypothesis thus provides a workable benchmark against which observed behavior may be contrasted. Its empirical performance to date is impressive, but better and more complete data may well indicate predictable exceptions to its implications that knowledgeable parties will attempt to exploit.

III. Implications

These strong findings favorable to the arbitrator exchangeability hypothesis have also been confirmed in further empirical work. (See Henry Farber and Max Bazerman, 1986; Bloom, 1986; and my paper with James Dow and Daniel Gallagher, 1986.) These findings have a number of implications for further research into the nature and design of arbitration systems. First, these results provide support for the pioneering theoretical approach used by Farber (1980) that explicitly recognizes the stochastic nature of arbitrator decisions. Second, the arbitrator exchangeability hypothesis provides a simple rationale for the study of alternative arbitration systems in laboratory experiments where the arbitrator decisions may conveniently be treated as a stable stochastic process. (See my paper with Janet Neelin and Matthew Spiegel, 1986.) Finally, the arbitrator exchangeability hypothesis provides a benchmark that can serve as a convenient null hypothesis in the search for predictable deviations to it that might be (profitably) exploited by the parties.

REFERENCES

Ashenfelter, Orley and Bloom, David, "Models of Arbitrator Behavior: Theory and Evidence," *American Economic Review*, March 1984, *74*, 111–25.

______, **Dow, James and Gallagher, Daniel,** "Arbitrator and Negotiation Behavior Under an Appelate System," mimeo., Industrial Relations Section, Princeton University, 1986.

______, **Neelin, Janet and Spiegel, Matthew,** "Experiments Comparing Alternative Arbitration Systems," mimeo., Industrial Relations Section, Princeton University, 1986.

Bloom, David E. "Empirical Models of Arbitrator Behavior under Conventional Arbitration," *Review of Economics and Statistics*, November 1986, *68*, 578–85.

______ **and Cavanagh, Christopher L.,** "An Analysis of the Selection of Arbitrators," *American Economic Review*, June 1986, *76*, 408–23.

Farber, Henry S., "An Analysis of Final-Offer Arbitration," *Journal of Conflict Resolution*, December 1980, *5*, 683–705.

______ **and Bazerman, Max H.,** "The General Basis of Arbitrator Behavior: An Empirical Analysis of Conventional and Final-Offer Arbitration," *Econometrica*, November 1986, *54*, 1503–28.

Klevorick, Alvin K., Rothschild, Michael and Winship, Christopher, "Information Processing and Jury Decisionmaking," *Journal of Public Economics*, April 1984, *23*, 245–79.

Joskow, Paul L., "Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets," *American Economic Review*, March 1987, *77*, 168–85.

Why is there Disagreement in Bargaining?

By HENRY S. FARBER AND MAX H. BAZERMAN*

One of the enduring puzzles in the analysis of bargaining is why there is ever disagreement in cases where agreement appears to be in the interests of both parties. In this study two types of arbitration schemes, conventional and final-offer, are described along with some evidence on relative rates of disagreement under the two schemes. A number of alternative explanations for disagreement are then outlined. The evidence on relative settlement rates in coventional and final-offer arbitration is used to evaluate the potential of these explanations to account for disagreement.

I. Some Background on Arbitration

Two types of arbitration are in wide use to settle disputes that arise in the negotiation of labor contracts among public sector employees. Conventional arbitration (CA), where the arbitrator is free to make any award he or she sees fit, was the first to be used. Final-offer arbitration (FOA), where the arbitrator is constrained to make an award that is equal to one or the other of the parties' last offers, was introduced in order to address the criticism of CA that arbitrator's tend to split the difference between the last offers of the parties.

Negotiated settlement rates are much higher under FOA than under CA (for example, Peter Feuille, 1975; David Grigsby and William Bigoness, 1982; Margaret Neale and Bazerman, 1983). However, the theoretical reasons for this are not clear. Thomas Kochan (1980) argues that the rationale for FOA are 1) that an arbitration procedure (or any dispute settlement procedure for that matter) is effective in encouraging negotiated settlements to the extent that it imposes costs on the parties in the event they fail to reach a negotiated settlement; and 2) that FOA is more effective in imposing these disagreement costs on the parties. Farber and Harry Katz (1979) argue that arbitration imposes costs on the parties largely due to the combination of risk aversion by the parties and their uncertainty regarding the behavior of the arbitrator. To the extent that the parties are risk averse, they will be willing to concede in negotiation from the expected arbitration award in order to avoid the risk of an unfavorable award.

The basis of the claim that CA is flawed is that if arbitrators split the difference their behavior is easily predictable. This results in no uncertainty and no cost of disagreement. In FOA no such splitting the difference is permitted so that the uncertainty is preserved along with the cost of disagreement (Farber, 1980a). However, Farber (1981) argues that arbitrators do not split the difference in CA because that would provide the parties with the incentive to make their last offers as extreme as possible. Such extreme behavior is not observed and our 1985 paper and David Bloom (1986) present evidence that arbitrators in CA pay attention to the facts of the case along with the offers. It may be that what seems like splitting-the-difference behavior is actually the parties presenting last offers that bracket the parties' expectation of the arbitrator's preferred settlement (Farber, 1981). The conclusion (Farber, 1980b) is that it is an empirical matter as to whether CA or FOA imposes higher costs of disagreement on the parties.

The manifestation of these costs of disagreement is a contract zone (a range of settlements that both parties prefer to disagreement). In one way or another, most theories of disagreement relate the likelihood of settlement to the existence or size of the

*Professor, Department of Economics, Massachusetts Institute of Technology, E52-252F, Cambridge, MA 02139 and Professor, Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60201, respectively. We thank Robert Gibbons for very useful comments on an earlier draft of this paper.

contract zone. Since the size of the contract zone in any particular case is a function both of expectations regarding arbitrator behavior and of the preferences of the parties, it is difficult to find evidence on the size of contract zones that is independent of settlement rates.

We use the calculations of contract zones in CA and FOA developed in our 1987 paper in the analysis that follows. These calculations are made assuming the parties have identical expectations regarding arbitrator behavior, and the results are called "identical-expectations contract zones" to distinguish them from de facto contract zones that might exist (or not exist) when expectations diverge. The bases of these calculations are empirical estimates of arbitrator behavior in CA and FOA we derived (1986) using data on decisions of professional arbitrators in hypothetical cases. The underlying theoretical models of negotiator and arbitrator behavior in CA and FOA are described in Farber (1981; 1980a). To close the model it was assumed that the parties have constant absolute risk-aversion utility functions and identical normal prior distributions on the arbitrator's underlying notion of an appropriate award. This distribution is the source of the parties uncertainty regarding the arbitration award. The clear conclusion from these calculations is that these contract zones are unambiguously larger in CA than in final-offer arbitration. However, the calculations also suggest that the Nash equilibrium last offers presented to the arbitrator are substantially closer together in FOA than in CA. These comparisons are robust with regard to wide ranges of values of risk aversions of the parties as well as wide ranges of values of the parameters of the conventional arbitrator behavior function (see our 1987 paper).

We turn now to a discussion of the various theories of disagreement with a set of three facts in hand. First, disagreement rates are lower in FOA than in CA. Second, identical-expectations contract zones are smaller in FOA than in CA. Third, the equilibrium last offers to be presented to the arbitrator are closer together in FOA than in CA.

II. Which Theories of Disagreement Fit the Facts?

A. *Divergent Expectations*

One prominent explanation for disagreement in bargaining is that the parties have divergent and relatively optimistic expectations regarding the ultimate outcome if they fail to agree. The most straightforward view of this model is that there will be agreement if and only if a contract zone exists but that relatively optimistic expectations can cause the nonexistence of a contract zone even where disagreement is costly. In the case where an arbitrator will render a decision if the parties fail to agree, relative optimism means that both parties expect to receive a relatively favorable decision from the arbitrator (for example, the union expects a higher wage award than the employer expects).

There is some empirical support for relative optimism in negotiations. Neale-Bazerman and Bazerman-Neale (1982), using data from a negotiation experiment, find that negotiators systematically overestimate the probability that they will be successful in arbitration. Thus, while relative optimism is not consistent with a simple equilibrium economic model, evidence suggests that negotiators systematically misperceive their environment in ways that could lead to disagreement.

Unless the costs imposed by uncertainty are sufficient to offset these divergent expectations completely, there will not be a contract zone and there will be no agreement. The identical-expectations contract zone, *because it is an indicator of the costs imposed by the uncertainty regarding the arbitration award*, is a direct measure of how robust the actual contract zone is to differences in expectations. Assuming that systematic differences in expectations are independent of the type of arbitration scheme, this divergent expectations model has the clear implication that larger identical-expectations contract zones will lead to a higher likelihood of agreement in actual cases where expectations may well differ.

The evidence is clearly not consistent with the divergent expectations theory. The identical-expectations contract zones we calculated elsewhere (1987) are substantially larger in CA than in FOA. This implies that contract zones under CA ought to be more robust to relative optimism in expectations. However, settlement rates are substantially higher under FOA. The conclusion is that divergent expectations are not a sufficient explanation for disagreement.

B. *Learning in Models with Asymmetric Information*

One class of models that has been suggested recently as an equilibrium explanation for disagreements in bargaining is based on the idea that there is asymmetric information held by party 1 that party 2 attempts to learn about by making offers that party 1 is free to accept or reject (for example, Drew Fudenberg and Jean Tirole, 1981; Joel Sobel and Ichiro Takahashi, 1983). A very simplified form of the argument is that firms have private information about their profitability that they cannot credibly transmit to the union. In a two-period model, the union formulates a first-period demand that the firm will accept if it is high profit (resulting in agreement) and reject if it is low profit (resulting in disagreement). This strategy is optimal from the union's point of view since it would like a high wage if possible. The firm can only make credible that fact that it is low profit by incurring the cost of disagreement. Hence, there will be disagreement some of the time. This theory has the clear implication that there will be less disagreement where the total costs of disagreement are higher.

Given that identical-expectations contract zone size is an indicator of total costs of disagreement, this simple theory predicts that there would be more disagreement where the identical-expectations contract zone is smaller. This is not consistent with the basic evidence that identical-expectations contract zones are larger and there is more disagreement in CA than in FOA. This suggests that learning in at least such simple models of asymmetric information is not a sufficient explanation of disagreement.

C. *Models of Commitment*

Another recent alternative class of models of disagreement has been developed by Vincent Crawford (1982) and is based on Thomas Schelling's (1956) model of commitment. The basic idea is that it may be advantageous in bargaining for the parties to commit to a position that would be very costly to disavow. Crawford develops a model where the potential for commitment by both parties can lead to disagreement as long as there is an element of irreversibility in the commitment and there is uncertainty about the strength of the parties' commitments. A sketch of the model is that commitment is reversible only at some uncertain cost and neither party knows this cost *ex ante*. In the first stage of bargaining in this model, the parties determine whether they will attempt commitment. In the second stage, the cost of backing down is revealed to each party but not to the other party. At this point, each party determines if they should back down, not knowing for certain whether the other party will back down.

Three classes of outcomes are possible. If both parties commit, there is disagreement. If only one party commits, that party gets a favorable settlement. Finally, if neither party commits, there is some solution concept that leads to agreement. The key is that there is a nonzero probability that both parties commit successfully, resulting in disagreement. Anything that increases the payoff to commitment will increase the probability of successful commitment and, hence, the probability of disagreement. Where the contract zone is large, a successful commitment may or may not have a larger payoff. Crawford argues that there is no clear prediction of the model regarding the extent to which the size of the contract zone affects the likelihood of commitment. However, he concludes that the conditions on the model required to predict unambiguously that larger contract zones lead to less commitment (and hence less disagreement) are not likely to be satisfied.

With regard to the evidence, this commitment theory has no implications for the relationship between contract zone size and the likelihood of agreement. Indeed, it is difficult to think of what evidence could be used to test this theory. Thus, while the commitment theory does not contradict the evidence, this is simply because of the lack of a clear prediction. The theory does not predict the strong pattern of evidence.

D. *Strategic Behavior in Arbitration Limits Concessions in Negotiations*

An explanation of disagreement that is specific to the arbitration setting is based on the structural feature of the arbitration process that arbitrators may receive information from the parties in hearings regarding the course of negotiations. In this context, it could be difficult for the parties to "retrench" from concessions made in bargaining in order to present the optimal offers to the arbitrators. Hoyt Wheeler's (1977) suggestion that arbitration procedures be modified so that the record of the negotiations are not admissible as evidence for the arbitrator is designed to address this problem. It is interesting to note that in the civil court system, which is perfectly analogous to labor arbitration (out-of-court settlement = negotiated settlement, trial outcome = arbitration award), offers to settle out of court are *not* admissible as evidence in a trial. This is precisely to avoid a reluctance to concede in attempts to reach a negotiated settlement.

This explanation suggests that where the equilibrium offers to be presented to the arbitrator are farther apart, the parties will be more reluctant to concede in bargaining, particularly if they are uncertain about whether a negotiated settlement is possible. In actual negotiations, each party may not know with certainty the other parties preferences or expectations about the arbitrator. In this situation, the parties are uncertain about whether any particular offer will be acceptable to the other side. In deciding whether to concede in making an offer, each party will weigh the expected gain from concession (the increase in the probability of agreement times the marginal gain from agreement) against the cost of making the concession (the decreased payoff from disagreement times the probability of disagreement). Where the optimal offers to be presented to the arbitrator are far apart, substantial concessions from these optimal offers are likely to be required.

Our 1987 finding of equilibrium offers in CA that are much farther apart than the equilibrium offers in FOA is clearly consistent with the explanation outlined in this section. Higher settlement rate in FOA may well be due to the convergence of the optimal offers in FOA combined with the structural features of the arbitration process that make it difficult to retrench to the optimal offers once concessions in negotiations beyond that point have been made.

E. *Salience of the Effect of the Offers on the Arbitration Award*

The final potential explanation of disagreement is based on the notion that the effect of the offers on the arbitration award is much more salient to the parties in FOA than in CA. That this is possible is obvious from the structures of the two procedures. In FOA, there is no escaping consideration of the direct effect that a party's offer will have on the arbitration award. In CA, the parties could well ignore the effects that their offers have on the awards and maintain extreme positions.

There is evidence that negotiators are not very good at working out the structure of the game that they are playing or considering the perspective of opponents and third parties (Bazerman and John Carroll, 1987). Bazerman-Neale and Neale-Bazerman show that FOA results in more concessionary behavior than CA precisely because its structure encourages each party to take the perspective of the other party, a cognitive ability which leads to more agreement. This is consistent with the evidence that settlement rates are higher under final-offer arbitration.

The "salience" theory does not deny the role that the contract zone plays, but it challenges the notion that the parties are able to understand all of the interrelationships in the game they are playing in CA. To the extent that this is the case, the predictions of a standard economic model, which assumes that the parties fully understand the game, are likely to be flawed. While there may be a contract zone, de jure (according to economic theory), there may not be one, de facto. This theory is one of a number of cognitive deficiency explanations of disagreement discussed in Bazerman and Carroll.

V. Conclusion

The evidence used in this study does not seem to be consistent with the divergent expectations theory or with simple models of learning with asymmetric information. Both of these theories imply that settlement rates ought to be higher in CA than in FOA because the identical-expectations contract zones are larger in CA reflecting larger costs of disagreement. However, the evidence is clear that settlement rates are higher under FOA. A model of commitment as a cause of disagreement is presented that has no clear prediction regarding the evidence. Finally, a pair of alternative explanations are presented that are consistent with the evidence that settlement rates are higher in FOA than in CA. These include reluctance to concede where the optimal offers for the arbitrator are far apart in fear that concessions could "come back to haunt them" in arbitration, and lack of salience of the role of the offers in CA. Overall, it seems likely that a single theory of disagreement applicable in all or even most settings does not exist. The structural features of the particular bargaining environment and dispute settlement mechanism in combination with the behavioral processes of decision making under uncertainty can have strong effects on bargaining behavior that must be considered.

REFERENCES

Bazerman, Max H. and Carroll, John S., "Negotiator Cognition," in B. M. Staw and L. L. Cummings, eds., *Research in Organizational Behavior*, Vol. IX, Greenwich: JAI Press, forthcoming 1987.

______ **and Farber, Henry S.,** "Arbitrator Decision Making: When are Final Offers Important?," *Industrial and Labor Relations Review*, October 1985, *40*, 76–89.

______ **and Neale, Margaret A.,** "Improving Negotiation Effectiveness under Final-Offer Arbitration: The Role of Selection and Training," *Journal of Applied Psychology*, December 1982, *67*, 543–48.

Bloom, David E., "Empirical Models of Arbitrator Behavior Under Conventional Arbitration," *Review of Economics and Statistics*, December 1986, *68*, 578–85.

Crawford, Vincent P., "A Theory of Disagreement in Bargaining," *Econometrica*, May 1982, *50*, 607–37.

Farber, Henry S., (1980a) "An Analysis of Final-Offer Arbitration," *Journal of Conflict Resolution*, December 1980, *24*, 683–705.

______, (1980b) "Does Final-Offer Arbitration Encourage Bargaining?," in *Proceedings of the Thirty-third Annual Meeting of the Industrial Relations Research Association*, Denver, September 1980, 219–26.

______, "Splitting-the-Difference in Interest Arbitration," *Industrial and Labor Relations Review*, April 1981, *35*, 70–77.

______ **and Bazerman, Max H.,** "The General Basis of Arbitrator Behavior: An Empirical Analysis of Conventional and Final-Offer Arbitration," *Econometrica*, November 1986, *54*, 1503–28.

______ **and** ______, "Divergent Expectations as a Cause of Disagreement in Bargaining: Evidence from a Comparison of Arbitration Schemes," NBER Working Paper No. 2139, January 1987.

______ **and Katz, Harry C.,** "Interest Arbitration, Outcomes, and the Incentive to Bargain." *Industrial and Labor Relations Review*, October 1979, *33*, 55–63.

Feuille, Peter, "Final Offer Arbitration and the Chilling Effect," *Industrial Relations*, October 1975, *14*, 302–10.

Fudenberg, Drew and Tirole, Jean, "Sequential Bargaining with Incomplete Information," *Review of Economic Studies*, November 1981, *50*, 221–48.

Grigsby, David and Bigoness, William "The Effects of Third Party Intervention on Pre-Intervention Bargaining Behavior," *Journal of Applied Psychology*, October

1982, *67*, 549–54.

Kochan, Thomas A., "Collective Bargaining and Organizational Behavior Research," in B. Staw and L. Cummings, eds., *Research in Organizational Behavior*, Vol. 2, Greenwich: JAI Press, 1980.

Neale, Margaret A. and Bazerman, Max H., "The Role of Perspective Taking Ability in Negotiating Under Different Forms of Arbitration," *Industrial and Labor Relations Review*, April 1983, *36*, 378–88.

Schelling, Thomas C., "An Essay on Bargaining," *American Economic Review*, June 1956, *46*, 281–306.

Sobel, Joel and Takahashi, Ichiro, "A Multi-Stage Model of Bargaining," *Review of Economic Studies*, July 1983, *50*, 411–26.

Wheeler, Hoyt N., "Closed Offer: An Alternative to Final-Offer Selection," *Industrial Relations*, October 1977, *16*, 298–305.

Negotiator Behavior under Arbitration

By DAVID E. BLOOM AND CHRISTOPHER L. CAVANAGH*

Bilateral bargaining lies at the heart of many important economic institutions. Even when there are substantial gains to trade, disputes are a natural and persistent element of bargaining situations. In order to economize on the cost of disputes, a number of mechanisms for preventing or resolving disputes have evolved. The public court system is perhaps the best known of these mechanisms, although private adjudication systems also exist. Moreover, because of increasing costs and congestion in the public court system, private mechanisms for adjudicating disputes have, in recent years, abounded in both number and scope.

One of the most popular private mechanisms for adjudicating disputes is arbitration. For example, in the area of labor-management relations, there were over 100,000 arbitration cases in the United States in 1985—about four times the number that took place just fifteen years earlier. Indeed, over the past two decades, nearly half the states in the United States have established interest arbitration mechanisms to resolve pay disputes involving groups of public employees. Arbitration is also widely used to resolve commercial disputes and, more recently, to resolve selected categories of civil disputes that might otherwise congest the public court system.

Although arbitration mechanisms can vary substantially in design, they all tend to involve a third-party to a dispute *determining* its resolution. Arbitration guarantees finality in the resolution of a dispute, generally in a timely and legitimate fashion that limits the erosion of a bargaining relationship that might result from an ongoing dispute.

Arbitration is a fascinating mechanism for economists to study. First, the possibilities for empirical analysis are often quite extraordinary. For example, in comparison to the public court system, there is more heterogeneity in the types of "arbitration experiments" available and less heterogeneity in the data (i.e., similar disputes are dealt with according to a wide variety of arbitration mechanisms). Moreover, the outcomes of wage and salary arbitration are relatively easy to quantify for purposes of empirical analysis and permit econometric models to build on much existing research on wage determination. Thus, the study of arbitration —a relatively simple mechanism for resolving disputes—may yield important insights into more complex legal mechanisms that are relatively difficult to model and to subject to empirical analysis.

Second, arbitration systems provide excellent settings for testing some of the most basic propositions of game theory. Arbitration is essentially a game with simple and well-specified rules in which a small number of players can be easily identified. The availability of "real-world" data on situations in which there are incentives to behave strategically provides remarkable opportunities for the analysis of game-theoretic behavior.

Third, insofar as arbitration mechanisms can be structured in different ways, studying arbitration might lead to improved designs. Indeed, theoretical work on arbitration has raised a number of policy-relevant issues whose resolution ultimately depends upon the results of empirical analysis.

The emerging empirical literature on the economics of arbitration has been primarily concerned with modeling the behavior of arbitrators under alternative forms of arbitration. It seems natural that the empirical literature turn next to consideration of the behavior of negotiators under arbitration (which typically depends critically on expectations about arbitrator behavior). Our chief purpose in the remainder of this article is to identify some of the issues that might sensi-

*Department of Economics, Harvard University, Cambridge, MA 02138. This research was partially supported by NSF Grant SES-8309148. We are indebted to Robert Gibbons for helpful discussions.

bly be raised by empirical economists studying arbitration from the point of view of the negotiating parties.

In a typical bargaining/arbitration situation, there are three key problems that negotiators must solve. First, they must decide whether to settle their dispute voluntarily or to proceed to arbitration. Second, they must adopt a strategy for selecting an arbitrator. Third, they must bargain prior to arbitration and, if they are unable to settle voluntarily, they must formulate final positions to advance in arbitration. The remainder of this article will discuss each of these three choice variables in turn.[1]

I. Negotiation vs. Arbitration

The American system of industrial relations exhibits a strong normative preference for resolving disputes without the aid of third parties. Thus, Carl Stevens' (1966) observation that arbitration mechanisms can be designed in ways that discourage their use was greeted enthusiastically by the proponents of arbitration. Briefly, Stevens likened arbitration to the strike as a mechanism for imposing costs of disagreement on bargainers and thereby promoting voluntary settlements. These costs are composed of the direct costs of using an arbitration mechanism and the indirect costs that are generated by the interplay of arbitral uncertainty and disputants' risk aversion.

Subsequent work by Henry Farber and Harry Katz (1979) formalized Stevens' notion by deriving expressions for a contract zone (i.e., a locus of potential settlement points, all of which are preferred by both parties to the settlement expected under arbitration). A contract zone may be generated either by the costs of arbitration, or by at least one bargainer having overly pessimistic expectations about an arbitrator's award. Conversely, a contract zone will tend to be small or nonexistent when at least one bargainer has overly optimistic expectations about an arbitrator's award.

These early analyses assumed that voluntary settlements would be reached whenever there was a contract zone. Thus, divergent and mutually inconsistent expectations seemed to be a key determinant of the resort to arbitration. More recent work has pointed out that the existence of a contract zone is necessary, but not sufficient, for arbitration to lead to a voluntary settlement because there may be substantial direct costs of negotiation as well as uncertainty about settlement points within the contract zone; it follows that wider contract zones do not, *ceteris paribus*, imply lower impasse probabilities (Bloom, 1981).

One of the most striking facts about arbitration requiring explanation is the substantial fraction of bargaining cases that end up being arbitrated: the steady-state rate of arbitration usage seems to vary between 15 and 30 percent in states with compulsory interest arbitration laws. In view of this fact, the theory that divergent expectations about arbitrator behavior explain the use of arbitration is less than satisfactory. It seems unlikely that bargainers would consistently be overly optimistic about the size of an arbitration award in the context of what is essentially a repeated game. Even in the context of one-shot bargaining, rational bargainers will tend to reconcile their prior expectations about an arbitrator's behavior in the negotiations leading up to arbitration (for example, see John Geanakopolos and Heracles Polemarchakis, 1982). Thus, alternative explanations are worth exploring.

Arbitration usage rates are not notably different in states with conventional arbitration provisions and those with final-offer arbitration provisions.[2] This fact tends to

[1] Provided arbitration is not compelled by law, the parties must jointly (and privately) decide whether they will use arbitration to resolve their disputes. The decision to arbitrate can be made either before or at the time a particular dispute arises. Although *ex ante* agreements to arbitrate disputes raise interesting economic issues, they are beyond the scope of this article.

[2] Under conventional arbitration, the arbitrator simply renders a decision that consists of his or her best judgement of a fair settlement. In contrast, the arbitrator is constrained to render a decision that consists of one or the other of the parties' final offers, without compromise, under final-offer arbitration.

contradict two early, but still influential, views about arbitration: 1) the view that final-offer arbitration would induce risk-averse bargainers to make concessions until their bargaining positions eventually overlapped, thereby eliminating the need for arbitration; and 2) the view that split-the-difference behavior on the part of conventional arbitrators (whether actual or perceived) would tend to "chill" negotiators from making concessions in the bargaining that precedes arbitration and thereby increase the probability of a dispute ending up in arbitration.

Another early theory held that arbitration would have a "narcotic effect" on bargainers, according to which bargainers would habitually avoid the arduous demands of bargaining in favor of arbitrated outcomes. Simple descriptive statistics do seem to indicate that there is substantial variation in arbitration usage across bargaining units. Whether this tendency is indeed evidence of a genuine narcotic effect is difficult to test because it requires establishing serial correlation in the use of arbitration—after controlling for heterogeneity across bargaining units.

Future analyses of the resort to arbitration might usefully build upon the notion that bargaining parties are not internally homogeneous entities with identical preferences. Bargainers often have constituencies whose future political support they desire. Insofar as arbitrators can be viewed as paid "scapegoats," the parties' final positions and their resort to arbitration might be modeled in a principal-agent framework. In this spirit, Vincent Crawford (1982b) develops a formal game-theoretic model of bargaining impasses based on the notion that parties may rationally commit to irreconciliable bargaining positions. The political pressure of the constituency makes concession after commitment costly and so can lead to impasse.

To date, there have been few attempts to empirically implement a structural model of the arbitrate/negotiate decision. Although some studies have been conducted using experimental data, further analysis, especially using available data from actual arbitration systems, is much needed.

II. The Selection of Arbitrators

One of the most salient differences between arbitration and the public court system is that disputants typically have some say in the appointment of arbitrators. The two leading arbitrator selection mechanisms operate by having an impartial agency (such as the American Arbitration Association) supply disputants with identical lists of an odd number of arbitrators (along with information on their backgrounds, fees, etc.). *Alternate strike* mechanisms work by having each party alternately cross a name off the list with the last name remaining becoming the appointed arbitrator. *Rank-veto* mechanisms work by having the parties each veto a prespecified number of arbitrators, rank the remaining ones in order of their preferences, and then refer the list back to the agency which then makes an appointment in accordance with those preferences. The opportunity for disputants to express their preferences for different arbitrators suggests an element of strategic interaction according to which negotiators may veto or give relatively unpreferred rankings to highly preferred arbitrators in an attempt to manipulate the selection process.

The analysis of arbitrator selection is interesting for several reasons. First, it yields direct information on the characteristics of arbitrators that negotiators find desirable. Second, measuring the similarity of union and employer preferences for individual arbitrators may yield insights into whether collective bargaining and arbitration are primarily institutions of conflict or cooperation. Third, it can provide information on the strategic sophistication of negotiators and on the importance of strategic interaction in bargaining. Finally, as we argue below, the process of arbitrator selection may be closely related to the use of arbitration.

The empirical literature on the selection of arbitrators is still in its infancy and much important work remains to be done. In a recent paper (1986a), we analyzed a set of data on actual union and employer rankings of different panels of arbitrators under a rank-veto mechanism. The results indicate that unions and employers have similar preferences: in favor of lawyers, more experienced arbitrators, and arbitrators who seem to have favored their side in the past. In addition, both sides exhibit strong preferences about having economists serve as arbitrators, with employers being in favor and unions being against. The analysis also tests whether the observed rankings data reveal the negotiators' true preferences over arbitrators. The results provide no support for the hypothesis of strategic misrepresentation of preferences by either side. Nonetheless, further empirical analysis of arbitrator selection is desirable, especially in the context of alternate strike mechanisms in which empirically falsifiable hypotheses about strategic behavior can be tested directly.

In another paper (1986b), we researched two aspects of the arbitrator selection phase of an arbitration system: 1) the strategies and outcomes of the selection "subgame" and 2) the impact of selection mechanisms on the bargaining environment. Our results on the selection subgame indicate that there is frequently no incentive to strategically misrepresent preferences—depending on the bargainers' preferences over arbitrators and on how much information they have about the other side's preferences. Thus, our earlier empirical results do not necessarily imply that negotiators are unsophisticated or irrational in their behavior. On the other hand, individually rational behavior in situations in which negotiators do have incentives to strategically misrepresent their preferences can result in inefficient selections.

This work also led us to conclude that arbitration is not necessarily best viewed in a purely static framework in which the size of the contract zone is fixed by an unchanging amount of arbitral uncertainty. We developed a model of the bargaining/arbitration process that has three distinct stages: 1) bargaining that takes place before the panel of prospective arbitrators is announced; 2) bargaining that takes place after the panel of prospective arbitrators is announced but before a particular arbitrator is selected; and 3) bargaining that takes place after a particular arbitrator is selected. Corresponding to each separate stage is a specific degree of uncertainty about the final resolution of the dispute. Furthermore, the degree of uncertainty tends to decrease as the parties move from one stage to the next. Although empirical analyses of the relationship between arbitrator selection mechanisms and the probability of impasse have yet to be conducted, we suspect that a dynamic mechanism that confronts bargainers with a varied set of bargaining environments is likely to provide them with better opportunities to reach agreement than a static mechanism that presents them with a single alternative. In practice, there is a substantial amount of voluntary settlement in each of the distinct stages of the bargaining/arbitration process.

III. Determination of Negotiators' Arbitration Positions

The earliest models of negotiators' arbitration positions presuppose that arbitrator preferences are imperfectly known to the disputants and depend solely on the exogenous facts of a dispute. In such a model, the mean of the distribution of arbitrator preferences becomes the focal point around which negotiators bargain, both in the negotiations that precede arbitration as well as in the arbitration process itself. If that focal point is different from the average settlement that would be negotiated in the absence of arbitration, the "option to arbitrate" will bias negotiated settlements. In addition, the negotiated settlements may not be Pareto efficient if there are multiple issues in dispute (see Crawford, 1979, 1982a).

Final-offer arbitration provides negotiators with an incentive to moderate their positions since less extreme positions presumably have higher probabilities of being selected by an arbitrator. However, because smaller payoffs are associated with more moderate positions, negotiators also have some incentive to adopt extreme positions. In the context of single-issue disputes, Nash final offers have the following properties: 1) the final offer of the more risk-averse negotiator will lie closer to the mean of the prior distribution of arbitrator preferences than the final offer of the less risk-averse negotiator; 2) increasing arbitral uncertainty by lowering the density of arbitrator preferences at the mean of the Nash pair of final offers causes those offers to diverge; and 3) even if both negotiators are risk neutral, arbitral uncertainty will cause their final offers to diverge (symmetrically) from the median of the arbitrator's preference distribution (see Farber, 1980).[3]

Under conventional arbitration, negotiators have literally no incentive to express a final position if arbitrator preferences are conditioned solely on exogenous background information. The fact that they almost always do suggests that the true model of arbitrator preferences may be somewhat different. An alternative (but equally polar) model of the behavior of conventional arbitrators is one in which they simply "split-the-difference" between the parties' final positions. The fact that conventional arbitration decisions typically lie near the average of the parties' final positions provides at least some empirical support for this view, although proper evaluation of this piece of data requires a model of negotiator behavior.

It seems obvious that the negotiators optimal final offers will diverge if arbitrators mechanically split-the-difference. An intermediate model, in which an arbitrator's preferred settlement depends on both the exogenous facts and the parties' final positions, would seem to be more plausible. This formulation suggests that arbitrators extract a useful signal about negotiator preferences from their final positions. If so, it follows that negotiators will have an incentive to communicate strategically to the arbitrator through their final offers.

The notion that negotiator final offers are attempts to "position the arbitrator" suggests that arbitrator and negotiator behavior should be modeled as a three-party game. It also suggests that econometric attempts to estimate the parameters of arbitrator preference functions by studying how arbitrator's behave when confronted with different sets of facts and final positions may be misspecified insofar as the final positions are endogenously determined. Put another way, it remains an open empirical question whether arbitrators should mainly be viewed as individuals who 1) impose on the negotiators their exogenous preferences, or 2) seek to learn about the disputants' preferences from the relationship between the facts and final positions in an attempt to search for outcomes that maximize the disputants' welfare. Robert Gibbons (1986) has begun the important task of modeling arbitrator behavior in the context of a three-party game of arbitration, although further work remains to be done.

In this connection, it is worth reflecting on the relation between the bargaining process that precedes arbitration and the arbitration game itself. Indeed, it is only reasonable to suppose that split-the-difference behavior on the part of arbitrators will "chill" negotiators from making concessions in the bargaining that precedes arbitration if one imagines that 1) negotiating concessions cannot be "taken back" in arbitration, and 2) arbitrators extract information from the parties' pre-arbitration behavior. In practice, both conditions are likely to be satisfied, suggesting a close coupling of behavior in negotiations and arbitration. Perhaps a small change in the design of arbitration mechanisms, in which arbitration hearings would be conducted without reference to the negotiations that preceded arbitration (i.e., de novo), would uncouple the two games and better serve the interests of both the negotiators and public policy.

Simple facts about the relation between negotiators' final positions under conventional and final-offer arbitration could be a good starting point for further theoretical work. Table 1 reports the average of employer and union final offers (EFO and UFO) in the two types of salary arbitration cases that took place in New Jersey in the years 1981–84. As the table makes clear, the parties' positions in conventional arbitration

[3] Although the negotiators' final offers are interdependent in this simple model, they may be independent of each other in a system of tri-offer arbitration such as the one that operates in Iowa. Orley Ashenfelter, James Dow, and Daniel Gallagher (1986) have done an interesting empirical analysis of the Iowa system that attempts to test a simple model of optimal negotiator behavior.

TABLE 1—UNION AND EMPLOYER FINAL SALARY OFFERS[a]

	1981	1982	1983	1984
Conventional Arbitration				
EFO	6.4	5.1	6.3	6.2
UFO	9.6	9.2	9.0	8.3
Final-Offer Arbitration				
EFO	7.2	7.2	6.5	6.6
UFO	9.0	9.3	8.3	7.8

Source: Authors' calculations based on New Jersey arbitration awards.

[a]EFO: Employer final offers; UFO: Union final offers; expressed as percent increases.

tend to lie outside the bounds of their positions in final-offer arbitration. This pattern is consistent with the predictions of the very simplest arbitration models according to which final-offer arbitration induces concessionary behavior by risk-averse bargainers while conventional arbitration chills the negotiation process that precedes arbitration. It remains to be seen whether more complete models of negotiator behavior under arbitration can further enrich our interpretation of these facts.

REFERENCES

Ashenfelter, Orley, Dow, James and Gallagher, Daniel, "Arbitrator and Negotiation Behavior under an Appelate System," unpublished, August 1986.

Bloom, David E., "Is Arbitration *Really* Compatible with Bargaining?," *Industrial Relations*, Fall 1981, *20*, 233-44.

______ **and Cavanagh, Christopher L.,** (1986a) "An Analysis of the Selection of Arbitrators," *American Economic Review*, June 1986, *76*, 408-22.

______ **and** ______, (1986b) "An Analysis of Alternative Mechanisms for Selecting Arbitrators," Harvard Institute of Economic Research Discussion Paper No. 1224, April 1986.

Crawford, Vincent P., "On Compulsory Arbitration Schemes," *Journal of Political Economy*, February 1979, *87*, 131-60.

______, (1982a) "Compulsory Arbitration, Arbitral Risk and Negotiated Settlements: A Case Study in Bargaining under Imperfect Information," *Review of Economic Studies*, January 1982, *49*, 69-82.

______, (1982b) "A Theory of Disagreement in Bargaining," *Econometrica*, May 1982, *50*, 607-37.

Farber, Henry S., "An Analysis of Final-Offer Arbitration," *Journal of Conflict Resolution*, December 1980, *35*, 683-705.

______ **and Katz, Harry C.,** "Interest Arbitration, Outcomes, and the Incentive to Bargain," *Industrial and Labor Relations Review*, October 1979, *33*, 55-63.

Geanakoplos, John and Polemarchakis, Heracles, "We Can't Disagree Forever," *Journal of Economic Theory*, October 1982, *28*, 192-200.

Gibbons, Robert, "Arbitration as a Signaling Game," unpublished, April 1986.

Stevens, Carl M., "Is Compulsory Arbitration Compatible with Bargaining?," *Industrial Relations*, February 1966, *5*, 38-52.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
NINETY-NINTH
ANNUAL
MEETING

NEW ORLEANS, LOUISIANA
DECEMBER 28–30, 1986

Minutes of the Annual Meeting
New Orleans, Louisiana
December 29, 1986

The ninety-ninth Annual Meeting of the American Economic Association was called to order by President Alice Rivlin at 5:35 P.M., December 29, 1986, in the Carondelet Ballroom of the New Orleans Marriott. The minutes of the meeting of December 29, 1985, were approved as published in the *American Economic Review, Papers and Proceedings* (May 1986, p. 421).

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Managing Editor of the *American Economic Review* (Orley Ashenfelter), Managing Editor of the *Journal of Economic Literature* (John Pencavel), Editor of the *Journal of Economic Perspectives* (Joseph Stiglitz) and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports published elsewhere in this issue. These had been distributed to members prior to the meeting.

Rivlin announced that the resolution submitted by W. Robert Brazelton and James Sturgeon had been withdrawn and would not be an item of business. The resolution read:

> Be it resolved that the *American Economic Review* and the *Journal of Economic Literature* allow all articles, invited or otherwise, to be subject to the publication of evaluations of them (such as subsequent or concurrent "Comments") subject only to the considerations of quality and relevance. This resolution assumes that no author published in either journal is not subject to a critical evaluation of his or her peers subject, once again, only to considerations of quality, relevance and published format style.

Rivlin then introduced Gary Becker, the 1987 President of the Association, to the audience. There being no further business, the meeting was adjourned.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in New York, New York, March 21, 1986.

The first meeting of the 1986 Executive Committee was called to order at 10:10 A.M. on March 31, 1986, in the Jolson-Cantor Room of the New York Marriott Marquis Hotel. Members present were Alice Rivlin (presiding), Orley Ashenfelter, Gary Becker, Alan Blinder, Peter Diamond, Rendigs Fels, Elton Hinshaw, Charles Kindleberger, Daniel McFadden, Janet Norwood, Mancur Olson, John Pencavel, Sherwin Rosen, and Charles Schultze. Leo Raskind was present as Counsel. Present for parts of the meeting were members of the Nominating Committee (W. Arthur Lewis, William Beeman, Padma Desai, Thomas Finn and Marjorie Honig) and Joseph Stiglitz, who gave a report.

Minutes. The minutes of the previous meeting (December 27, 1985) were approved without correction.

Report of the Secretary (Hinshaw). The 1986 annual meeting will be held in New Orleans, December 28–30. The schedule for subsequent meetings is Chicago (1987), New York (1988), and Atlanta (1989). Registration for the 1985 meeting in New York totaled 7,349. Thirty-eight other associations, societies, and organizations met with us, 355 scholarly sessions were held, and 94 other events (cocktail parties, committee meetings, lunches, etc.) were scheduled. The last time we met in New York (1982), total registration was 6,715, 34 other groups met with us, 328 scholarly sessions were held, and 94 other events scheduled.

Peter Navarro and Richard Carson (professors at the University of California, San Diego) had written the Secretary requesting that a research project on the academic job market they propose be conducted under the auspices of the AEA. The Executive Committee decided not to act as sponsor but encouraged the researchers to undertake the project.

Report of the Managing Editor of the American Economic Review (Ashenfelter). The June 1986 issue will contain the first articles selected by the current editorial procedures. Previous issues were the backlog inherited from Clower. Ashenfelter alerted the Executive Committee to his (and other editors of scholarly journals) growing concern about the nature of the data being used in some empirical research; many of the published results cannot be replicated. Because he expects to publish empirical studies, he is considering instituting a procedure for checking data sources against results.

Report of the Managing Editor of the Journal of Economic Literature (Pencavel). Pencavel reported that the transition from Abramovitz had gone smoothly and was now complete. Abramovitz continues as an Associate Editor. Pencavel noted that royalties from DIALOG were increasing significantly and that consideration was being given to adding working papers to the DIALOG system.

1986 Program (Becker). Becker reported that he and his Program Committee had organized about 60 sessions. About 10 of them would be devoted to the topic of the relationship between economics and other fields. Judge Richard Posner had agreed to give the Ely Lecture.

New Journal (Stiglitz). As requested at the last meeting of the Executive Committee, Stiglitz presented a more detailed budget and set of goals and procedures for the proposed new journal. The new journal is intended to serve the audience of professional economists which is not well served by a research journal. It would be directed to the consumers of research articles rather than the producers. Its goal would be to make the latest research more accessible to the average economist.

The editorial structure would include a co-editor and a professional writer-editor; 10–12 associate editors (probably textbook writers) to cover subdisciplines; an advisory board to help shape the journal and give feedback on the extent to which the journal was meeting the objectives; and a review panel of persons from the intended audience.

Stiglitz thought that most articles would be solicited rather than coming in over the transom. Articles would be reviewed for general interest, content, and style as opposed to originality of contribution. Emphasis will be on exposition.

It was VOTED to establish a journal of the type described, subsidize it up to a maximum of $500,000, appoint a Committee to develop an evaluation procedure, evaluate it after two years of issues, and send the first year's issues to all members and subscribers without charge. It was understood that Stiglitz would edit the new journal for three to five years and that the journal was expected to become "self-supporting."

Nominating Committee (Lewis). The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, choose Robert Eisner as the nominee for President-elect. Lewis reported the following nominees for other offices: for Vice-President (two to be chosen), Ann Friedlaender, Richard Cooper, Robert Lucas Jr., and Burton Weisbrod; for members of the Executive Committee (two to be chosen), Bernard Anderson, Judith Thornton, Rudolf Penner, and Robert Barro.

Other Business. It was VOTED to adopt the following policy of disclosure for the Association's journals: Authors are expected to reveal the sources of any financial or research support received in connection with the preparation of their article.

It was VOTED to adopt the following policy on advertising in the Association's publications: The Association will not knowingly publish paid or unpaid advertising or notices for products, services, or positions that are made available in a fashion that discriminates on the basis of race, color, religion, gender, sexual preference, or physical handicap.

It was VOTED to appoint the Secretary to another three-year term which would end on December 31, 1990. It was agreed that he would also serve as Treasurer for one year after Fels's resignation as Treasurer becomes effective (December 31, 1987).

It was agreed that the President would appoint a Committee to coordinate Association efforts on government statistics and improvement thereof.

It was decided not to pursue the development of an Association-sponsored insurance plan for excess major medical expenses.

Report of the Treasurer (Fels). Rivlin announced that Fels had resigned as of the end of his current term. After a general expression of gratitude for him and his work by members of the Executive Committee, he reported that although the Association has been having operating losses, investment gains, augmented by rising stock prices, have produced surpluses aggregating $532 thousand in the last two years. The ratio of the net worth at the end of 1985 to budgeted expenses for 1986 is 1.75, far above the ratio of 1.0 deemed more than adequate for safety purposes.

The investment gains include capital gains, whether realized or not, adjusted for inflation. (Capital gains from equities are recognized over a three-year period.) Some time ago the Budget Committee recommended abandoning the investment-income formula now used in favor of counting as investment income 4 percent of the market value of the portfolio. The Executive Committee has not yet taken action on this recommendation. In the past, the 4 percent formula would have shown total investment gains over the years about the same as the formula actually used but there would have been less variability. If the 4 percent formula had been in effect in 1984 and 1985, the aggregate surpluses for the two years would have been only $68 thousand instead of $532 thousand. This points up the likelihood that investment gains in the future will be lower than those shown by the income statements for 1984 and 1985.

The decision to inaugurate a new journal with start-up costs of half a million dollars, reducing investment income and increasing operating costs, suggests that dues and subscription prices be increased effective January 1, 1987. He proposed increasing the base rate of dues from $37.50 to $38.50 and the price of subscriptions from $105 to $110. Such increases can be justified both by the rate of inflation and by the expectation that in 1987 one or two issues of the new journal will be distributed to all members and subscribers without charge and without reduction in other benefits.

At its meeting in December, the Executive Committee should consider how to finance the new journal in the long run. A decision should be made no later than March 1987. Among the options are: (1) Increase dues and subscription prices; distribute the new journal to all members and subscribers. (2) Finance the new journal entirely by subscriptions separate from the *AER* and the *JEL*, subjecting it to a market test. (3) Unbundle, giving members and subscribers their choice of two of the three journals, with a low extra charge (a little over marginal cost) for all three. (4) Another form of unbundling with members getting one journal of their choice with extra charges for the others. (Dues would be lower under 4 than under 3.)

It was VOTED to approve the recommended increases in dues and subscriptions and the 1986 budget.

The meeting adjourned at 3:20 P.M.

Minutes of the Meeting of the Executive Committee in New Orleans, Louisiana, December 27, 1986.

The second meeting of the 1986 Executive Committee was called to order at 10:05 A.M. on December 27, 1986, in the St. Charles Room of the New Orleans Marriott Hotel, New Orleans, Louisiana. Members present were Alice Rivlin (presiding), Orley Ashenfelter, Gary S. Becker, Alan S. Blinder, Peter A. Diamond, Rendigs Fels, Victor R. Fuchs, C. Elton Hinshaw, Charles P. Kindleberger, Daniel McFadden, Janet L. Norwood, Mancur Olson, Jr., John Pencavel, Sherwin Rosen, Thomas J. Sargent, and Charles L. Schultze. Also present as guests for all or part of the meeting were Robert A. Barro, Nina Cornell, Robert Eisner, Michael McCarthy, Ronald L. Oaxaca, Leo Raskind (AEA Counsel), and Isabel Sawhill.

President Rivlin welcomed Robert Eisner as the new President-elect and thanked those members whose terms were expiring for their dedicated service (Diamond, Fuchs, Norwood, Olson, and Schultze). She then called for consideration of the minutes of the previous meeting. They were approved as written.

Report of the Secretary (Hinshaw). The Secretary reminded the Committee that the 1987 annual meeting will be held in Chicago, December 28–30, the 1988 meeting in New York and the 1989 one in Atlanta. Washington, Boston, and San Francisco are being considered as possibilities for 1990. He also stated that, although the final count is not yet available, this New Orleans meeting promised to be the largest non-East Coast meeting in AEA history.

If the Editor of the new journal, *Journal of Economic Perspectives* (*JEP*), is to have the same status as the Managing Editors of the *American Economic Review* (*AER*) and the *Journal of Economic Literature* (*JEL*), three paragraphs of the Association's bylaws will need amending. It was VOTED to approve the changes recommended by the Secretary and submit them to the members in a mail ballot. The amended bylaws would read as follows:

> *Section III, Paragraph* 2. The Association shall have the following officers who shall be appointed by the Executive Committee: a Secretary, a Treasurer, the Editors of its scholarly journals, and a Counsel. The terms of office of each of these officers shall be three calendar years.
> *Section III, Paragraph* 3. The Executive Committee shall consist of the President, the President-elect, two Vice-Presidents, the Secretary, the Treasurer, the Editors, the two ex-Presidents who have last held office, and six elected members, provided the Secretary, the Treasurer, and the Editors shall not be entitled to vote in the Executive Committee's meetings.
> *Section IV, Paragraph* 7. The Editors shall, with the advice and consent of the Executive Committee, appoint members of Editorial Boards to assist them. The editors shall be *ex officiis* members and chairpersons of their respective Boards, which shall have charge of the publications.

A publisher had contacted the Secretary and proposed an agreement to publish some new volumes of readings and surveys. Discussion of the proposal indicated that participation in such a project was no longer necessary to demonstrate the commercial viability of "books of readings." It was VOTED not to pursue the proposal.

The Secretary asked for guidance in interpreting a motion passed at the March 21, 1986, meeting: "The Association will not knowingly publish paid or unpaid advertising or notices for products, services or positions that are made available in a fashion that discriminates on the basis of race, color, religion, gender, sexual preference, or physical handicap." He had understood that the Executive Committee had meant to allow private, sectarian colleges and universities to express a preference for members of their own faith but the motion, if read literally, prohibits that. It was agreed that the AEA should not accept advertisements containing language contrary to the motion. Preferences, based on categories mentioned in the motion, even if legal, may not be expressed in AEA publications.

Report of the Editor of the American Economic Review (Ashenfelter). Ashenfelter briefly reviewed his written report (see elsewhere in this issue). Acting on his recommendation, it was VOTED to approve the appointment of George Akerlof, Jo Anna Gray, Robert Porter, Richard Roll, and Kenneth Singleton to the *AER* Board of Editors.

He noted that the number of submissions was now approaching 1,000 annually, many of which were not finished manuscripts, and the submission fee in constant dollars was below what it was in 1972. Ashenfelter proposed an increase in the nominal submission fee to $50 for members and $100 for nonmembers to attempt to discourage the less-than-serious author. He further proposed instituting a $35 payment to referees who do their work promptly. The increased submission fees should just about equal the prompt-referee payments. It was pointed out that people would expect more refereeing to accompany the higher submission fee and that he might consider refunding the fee if one of the editors rejected a paper without sending it to outside referees. It was VOTED to raise the submission fee to $50 for members and $100 for nonmembers and to authorize a $35 payment to referees who do their critiques in a timely fashion.

Ashenfelter then raised the issue of double-blind refereeing (i.e., the author does not know who the referee is and the referee does not know who the author is). He and his Co-editors were divided on the value of blind refereeing. However, when George Borts was editor of the *AER*, he had used such a procedure and collected data on the results. The data had never been analyzed. Ashenfelter said that he wanted to have a colleague study the information available and would report the results to the Executive Committee. He was urged to report on the study at the March meeting when the issue would be considered again.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel briefly reviewed his written report (see elsewhere in this issue). The increase in the size of the journal has been caused primarily by the growth in the bibliographic section, reflecting the increase in the number and size of periodicals covered. If such growth continues, a major department of the journal may have to be eliminated to control costs. Two developments may indicate the proper direction for change: (1) If the new journal is successful, some of the articles published there might have sufficient overlap with the articles department of the *JEL* to allow a reduction in *JEL* space devoted to articles. (2) DIALOG, an on-line bibliographic retrieval service, may allow a reduction in space devoted to the bibliographic department. He foresees a change in the nature and character of the *JEL* over the next 5 to 10 years as he tries to cope with the increasing growth of periodicals.

Report of the Editor of the Journal of Economic Perspectives (Stiglitz). Stiglitz briefly reviewed his written report (see elsewhere in this journal). It was VOTED to approve his recommendations for members of the Board of Editors: Henry J. Aaron, Stanley Fisher, Paul R. Krugman, Edward P. Lazear, Mark J. Machina, Charles F. Manski, Donald N. McCloskey, Bernard Saffran, Steven C. Salop, Lawrence H. Summers, Hal R. Varian, and Janet L. Yellen. It was VOTED to approve the appointment of Carl Shapiro as Co-editor. Stiglitz and Ashenfelter recommended the transfer of the "Notes" section of the *AER* to the *JEP*. This was approved. Stiglitz's request that he be allowed to seek outside financial support for symposia was approved.

In response to a question about the potential overlap of *JEP* articles with *JEL* articles, Stiglitz said that there is a clear distinction between the two. *JEP* articles will not be literature surveys. They will offer perspectives by several authors on the same topic with explanations of how economics provides perspective on policy problems. The first issue is expected to be published this summer.

Report of the Director of Job Openings for Economists (Hinshaw). Hinshaw referred the Committee to his written report published elsewhere in this journal. There were no questions or comments.

1987 Program (Eisner). Eisner, President-elect and Program Chair for the 1987 meetings stated the theme would be "The Challenge of Full Employment." This may indeed be construed broadly and comprehend a wide body of work—theoretical, economic, historical, and policy-oriented. It need not discourage those who feel no such challenge or wish to demonstrate once more that all unemployment is "natural" or voluntary. It may encompass papers on information theory, human capital, and the nature (or incompleteness) of current and futures markets for labor, commodities, and financial claims. It may encompass issues of discrimination with regard to sex, race, and age. And it certainly may extend as well to questions of fiscal and monetary policy and theory.

Affirmative action with regard to sessions relating more or less to the "theme" will of course in no way imply exclusion of contributions reflecting the broad range of current interests among economists—and the broad range of economic issues facing the nation.

Report of Representatives to the Consortium of Social Science Associations (Cornell). Cornell reminded the group of COSSA's purpose—to encourage federal funding of basic research in the social sciences. Its primary focus is on the National Science Foundation budget with a secondary emphasis on the national institutes of Health. She judged that COSSA was beginning to look for other things to do than just plead the case for basic research and sought advice about what stance she should take toward potential, new directions. It was the consensus that any departure from the original purpose of COSSA should be done with great care and slowly, if at all. Further, it was opined that nothing could be more important than keeping federal funding of research within the peer review system. COSSA should be urged to take a strong stand on the issue.

Committee on the Status of Minorities in the Economics Profession (Oaxaca and McCarthy). Oaxaca, Chair of the Committee, reported that five two-year and six one-year fellowships for graduate work were offered for this coming academic year. Nineteen students had been nominated from 17 universities. He has proposed to the Rockefeller Foundation that the stipend be raised to $700 a month to make the stipend consistent with those provided by Ford Foundation and National Science Foundation fellowships. Since the Rockefeller grant expires at the end of 1987, he is seeking support for the fellowship program from other potential sponsors.

McCarthy, Director of the AEA Summer Program for Minority Students at Temple University, reported that 28 students completed the 1986 program—20 were blacks, 7 were Hispanics, and one a Native American. It was judged that, with the proper choice of institution and degree program, more than half of the students could be successful in graduate school. Funding for the 1987 program is in place; some $30–40 thousand will be needed for the summer of 1988.

Comments on McCarthy's report raised questions about how to evaluate the program in the absence of a control group, how effective was the selection process in predicting success in the program, and whether to define "minorities" (currently defined to be blacks, Hispanics, and Native Americans) to cover other groups that are underrepresented in the profession. No action was taken.

Committee on the Status of Women in the Economics Profession (Sawhill). Sawhill, Chair of the Committee, reviewed CSWEP activities during the past year and discussed her written report published elsewhere in this journal. Discussion of her report centered on the issue of blind-refereeing. The Committee believes that double-blind refereeing is strongly to be preferred as a matter of principle. It is likely to be perceived as fairer by women and less established members of the profession as well. It was decided to consider the issue again at the March meeting of the Executive Committee.

Investment Income Formula (Fels). Fels reported once again on the Budget Committee's recommendation that the formula for calculating investment income be changed. Under the current formula, real capital gains (or losses), whether realized or not, are recognized in three annual installments. The proposed change is to recognize a fixed percent of investible assets as income. The chief argument in favor of a fixed-percent rule is simplicity. A second argument is that recognized income would be less volatile, with a third argument being greater predictability. It was VOTED to adopt the fixed-rate method and use 5 percent as the rate for calculating income.

Other Business. A. H. Studenmund had proposed that the AEA sponsor a study of "article popularity" that uses reader responses rather than reference citations to attempt to measure the kinds of articles that are actually read and used. It was VOTED not to sponsor the project.

Julian L. Simon had proposed that the *AER* should no longer be the official journal of the AEA. Instead, there should be a system whereby members can buy a subscription to any journal they wish as part of membership, that is, some kind of a voucher system. The AEA might make bulk purchases of other journals and then retail them to members. The AEA would essentially become a journal broker. It was decided to postpone consideration of the proposal until a decision was made about permanent financing for the new AEA publication, *Journal of Economic Perspectives*. At that time one of the possibilities to be considered will be the unbundling of all AEA publications, that is, giving the members an option as to which journals they wish to receive.

W. Robert Brazelton and James Sturgeon had submitted a resolution for consideration at the annual business meeting:

> Be it resolved that the *American Economic Review* and the *Journal of Economic Literature* allow all articles, invited or otherwise, to be subject to the publication of evaluations of them (such as subsequent or concurrent "Comments") subject only to the considerations of quality and relevance. This resolution assumes that no author published in either journal if not subject to a critical evaluation of his or her peers subject, once again, only to considerations of quality, relevance and published format style.

It was agreed that the resolution was not needed and redundant, given our bylaws. Comments are published based on their quality, relevance, and importance. Being "right" is not necessarily sufficient. It was agreed that a member of the Executive Committee would discuss the issue with the author of the resolution and seek its withdrawal.

Report of the Treasurer (Fels). Fels reported that when the final results for 1986 become available, there probably will be a small surplus. The proposed budget for 1987 shows an operating loss of $413 thousand. Inasmuch as the net worth of the Association exceeds what is needed for safety, the projected deficit should not be considered dangerous to the Association's financial health. See his written report and the 1987 proposed budget elsewhere in this journal. It was VOTED to adopt the 1987 budget as proposed.

There being no further business to consider, the meeting adjourned at 5:10 P.M.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1986

Annual Meetings. In 1987, the annual meeting will be held in Chicago on December 28–30. The schedule for subsequent meetings is New York in 1988 and Atlanta in 1989. Each of these meetings is scheduled for December 28–30 and each will have a Placement Service, which will open for business one day earlier (December 27) than the meetings.

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of W. Arthur Lewis, Chair, William J. Beeman, William J. Boyes, Padma Desai, Nicholas Filippello, Thomas J. Finn, and Marjorie Honig submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-Elect
Robert Eisner

Vice-President	*Executive Committee*
Richard N. Cooper	Bernard E. Anderson
Ann F. Friedlaender	Robert J. Barro
Robert E. Lucas, Jr.	Rudolph G. Penner
Burton A. Weisbrod	Judith Thornton

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots. I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)
Robert Eisner
Vice-Presidents (for a term of one year)
Ann F. Friedlaender
Robert E. Lucas, Jr.
Executive Committee (for a term of three years)
Robert J. Barro
Judith Thornton

In addition, I have the following information:

Number of legal ballots	4,958
Number of invalid envelopes	162
Number of envelopes received after October 1	63
Number of envelopes returned	5,183

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of year)

	1984	1985	1986
Class of Membership			
Annual	16,612	17,602	17,148
Junior	1,932	1,670	1,632
Life	370	359	358
Honorary	29	30	33
Family	422	472	457
Complimentary	521	473	478
Total Members	19,886	20,606	20,106
Subscribers	5,846	5,852	5,846
Total Members and Subscribers	25,732	26,458	25,952

Membership. The total number of members and subscribers is shown in Table 1. The total has fluctuated between 25,000 and 26,500 since 1975 when it reached an all-time high of 26,787.

National Registry. The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 280 in 1986, compared to 381 in 1985. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

AEA Staff. Mary Winer, Kimberly Adair, Norma Ayres, Violet Sikes, and Jacquelyn Woods handle the day-to-day operations of the Association; Marlene Keefer organizes the operation of the annual meeting. Their dedication and efficiency make the job of the Secretary tolerable. I wish to express my great gratitude for the excellent work they continue to do.

Committees and Representatives. Listed below are those who served the Association during 1986 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them all for their services.

Budget Committee (3-year terms)
Rendigs Fels, *Chair*
Janet L. Norwood (1986)
Daniel McFadden (1987)
Sherwin Rosen (1988)
Alice M. Rivlin, *ex officio*
Gary S. Becker, *ex officio*
C. Elton Hinshaw

Census Advisory Committee
Morris A. Adelman, *Chair* (1987)
Rosanne E. Cole (1987)
Ben E. Laden (1987)
Victor Zarnowitz (1987)
Timothy F. Bresnahan (1988)
Robert J. Genetski (1988)
Darwin Johnson (1988)
Joel Popkin (1988)
Margaret C. Simms (1988)

Committee on Economic Education
W. Lee Hansen, *Chair* (1987)
Marianne A. Ferber (1986)
Michael K. Salemi (1987)
William B. Walstad (1987)
Bruce R. Dalgaard (1988)
Kalman Goldberg (1988)
Phillip Saunders, Jr. (1988)
Rendigs Fels, *ex officio*

Economics Institute Policy and Advisory Board
Edwin S. Mills, *Chair* (1986)
John E. Moroney (1986)
W. Lee Hansen (1987)
Dwight Perkins (1987)
Lance E. Davis (1988)
Stefan H. Robock (1988)
Joseph Havlicek, Jr. (1989)
Teh-wei Hu (1989)

Committee on Federal Economic Statistics
F. Thomas Juster, *Chair* (1988)
Barry P. Bosworth (1988)
John Cogan (1988)
Rosanne Cole (1988)
Ivan Felligi (1988)
Lyle E. Gramely (1988)
Zvi Griliches (1988)
William A. Morrill (1988)
Eugene Smolensky (1988)
Robert Solomon (1988)
John Wilson (1988)
Nina W. Cornell, *ex officio*
Marilyn Moon, *ex officio*
Joel Popkin, *ex officio*
Alice M. Rivlin, *ex officio*

Finance Committee
Rendigs Fels, *Chair*
Robert J. Genetski (1986)
Robert G. Dederick (1987)
Robert Eisner (1988)
C. Elton Hinshaw, *ex officio*

Committee on Honorary Members
Richard A. Musgrave, *Chair* (1986)
Hal R. Varian (1986)
Richard E. Caves (1988)
Franco Modigliani (1988)
J. Carter Murphy (1990)
Gordon C. Winston (1990)

Committee on Honors and Awards
Oliver E. Williamson, *Chair* (1991)
Robert Eisner (1987)
William Vickrey (1987)
James J. Heckman (1989)
Richard R. Nelson (1989)
Dale W. Jorgenson (1991)

1986 Nominating Committee
W. Arthur Lewis, *Chair*
Padma Desai
Thomas J. Finn
Marjorie Honig
A. Nicholas Filippello
William J. Beeman
William J. Boyes

Committee on Political Discrimination
Robert J. Lampman, *Chair* (1986)
Herbert Gintis (1986)
Richard R. Nelson (1986)
Benjamin J. Cohen (1987)
Clark W. Reynolds (1987)
Lester C. Thurow (1988)

Committee on the Status of Minority Groups in the Economics Profession
Ronald L. Oaxaca, *Chair* (1988)
Bernard E. Anderson (1987)
William A. Darity (1987)
Glenn Loury (1987)
Rhonda Williams (1987)
George Borjas (1988)
Vernon Dixon (1988)
Clifford E. Reid (1988)
Margaret Simms (1988)

Committee on the Status of Women in the Economics Profession
Isabel V. Sawhill, *Chair* (1987)
Lourdes Beneria (1986)
Bernadette Chachere (1986)
Mary Fish (1986)
Sharon B. Megdal (1986)
Michelle J. White (1986)
Karen Davis (1987)
Helen Junz (1987)
Beth E. Allen (1988)
Alan E. Fechter (1988)
Nancy Gordon (1988)
Katharine C. Lyall (1988)
Alice M. Rivlin, *ex officio*
Joan G. Haworth, Membership Secretary

AEA/SSRC Joint Committee on U.S.–China Exchanges
Gregory C. Chow, *Chair*
Kenneth Arrow
Lawrence R. Klein
Theodore W. Schultz

Committee on U.S.–Soviet Exchange
Franklyn D. Holzman, *Chair* (1987)
Jennifer R. Reinganum (1986)
Lloyd G. Reynolds (1986)
Abram Bergson (1987)
Joseph A. Pechman (1988)
Richard N. Rosett (1988)

Council and Other Representatives

American Association for the Advancement of Science, Section K, Social Economics and Political Sciences
Adam Rose (1988)

American Association for the Advancement of Slavic Studies
Joseph Brada (1989)

American Council of Learned Societies
C. Elton Hinshaw (1990)

Review Board of the American Statistical Association-Bureau of Census
Zvi Griliches

Review Board of the American Statistical Association (ASA)/Bureau of Labor Statistics (BLS)— Research Fellowship and Associate Program
Robert Pollak

Consortium of Social Science Associations (COSSA)
Nina Cornell (1988)
C. Elton Hinshaw

Council of Professional Associations on Federal Statistics (COPAFS)
Marilyn Moon (1988)
Joel Popkin (1988)

International Economic Association
Kenneth Arrow (1990)
C. Elton Hinshaw

Policy Board of the Journal of Consumer Research
Louis L. Wilde (1988)

National Bureau of Economic Research
David A. Kendrick (1987)

National Council for Social Studies
W. Lee Hansen

Social Science Research Council
Hugh T. Patrick (1987)

Representatives of the Association on Various Occasions—1986

Inaugurations
Peter Diamandopoulos, Adelphi University
Richard F. Dowd
Jay L. Kesler, Taylor University
Stanley R. Keil
Paige E. Mulhollan, Wright State University
Janet C. Goulet

C. Elton Hinshaw, *Secretary*

Report of the Treasurer for the Year Ending December 31, 1986

When the final results for 1986 become available, there probably will be a small surplus rather than the deficit that was expected last spring. Investment gains through the first nine months of 1986 exceeded what had been expected for the full year by $163 thousand.

The budget for 1987 approved by the Executive Committee on December 27, 1986, shows an operating loss of $413 thousand. If investment income turns out to be 4 percent of the value of the Association's portfolio on September 30, 1986, nearly half of the operating loss will be offset by investment gains. (See Table 1.)

Inasmuch as the net worth of the Association exceeds what is needed for safety, the projected deficit can be absorbed easily. Nevertheless, action will need to be taken eventually to increase revenues. Operating

TABLE 1—1987 BUDGET, AMERICAN ECONOMIC ASSOCIATION
(Thousands of dollars)

	First Nine Months		Full Year		
	(Unaudited)		Actual	Budgeted	
	1985	1986	1985	1986	1987
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	$614	$647	$831	$850	$870
Nonmember subscriptions	456	483	614	638	680
Subtotal	1,070	1,130	1,445	1,488	1,550
Subscriptions, *Job Openings for Economists*	20	21	30	30	30
Advertising	79	95	108	108	130
Sale of *Index of Economic Articles*	47	9	56	120	150
Sales of copies, republications, handbooks	22	30	27	32	27
Sale of mailing list	25	30	46	46	46
Annual meeting	16	37	16	16	16
Sundry	45	48	64	64	64
Total Operating Revenue	**1,323**	**1,400**	**1,792**	**1,904**	**2,013**
PUBLICATION EXPENSES					
American Economic Review	481	457	610	597	642
Journal of Economic Literature	569	582	780	758	852
Directory	41	53	50	70	70
Job Openings for Economists	35	36	54	55	57
Index of Economic Articles	41	6	45	100	75
Journal of Economic Perspectives		36		85	278
Subtotal	1,167	1,170	1,539	1,665	1,974
OPERATING AND ADMINISTRATIVE EXPENSES					
General and Administrative	187	220	325	335	350
Committees	16	27	42	50	50
Support of other organizations	48	49	50	56	52
Subtotal	251	296	417	441	452
Total Expenses	**1,418**	**1,466**	**1,956**	**2,106**	**2,426**
OPERATING GAIN (LOSS)	(95)	(66)	(164)	(202)	(413)
INVESTMENT GAIN (LOSS)	171	336	449	173	182
SURPLUS (DEFICIT)	76	270	285	(29)	(231)
Ratio, Net Worth to Annual Expenses			**1.75**		

costs will continue to rise. Deficits will reduce the portfolio and the income derived from it, increasing the deficits. The margin of safety provided by the net worth will diminish.

Audited financial statements for 1986 will be published in the June issue of the *American Economic Review*.

I am deeply indebted to our accountant, Norma Ayres, who over the years has done consistently outstanding work.

RENDIGS FELS, *Treasurer*

Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, Illinois at noon on December 15, 1986. Robert Dederick, Robert Eisner, Rendigs Fels (chairman), and Robert Genetski were present as members of the Committee. Harvey Hirschhorn, Robert McNeill, and James Weiss represented Stein Roe & Farnham, the Investment Counsel of the Association. C. Elton Hinshaw, Secretary and Treasurer-elect of the Association, was present as a guest.

I reported that the Association would need about half a million dollars from the portfolio between April 1 and September 30, 1987, and would continue to reduce the portfolio thereafter until the net worth of the Association was approximately equal to its annual expenditures.

Mr. Hirschhorn presented a forecast of economic conditions for the next two years. Mr. McNeill presented a written report that included minutes of selected past meetings, charts and tables relating to market performance, and data on the portfolio of the Association. During the past two years, the return on the total account of the Association was 47.2 percent. On the equities part of the portfolio, the return was 53.3 percent. During the same period the return on Standard & Poor's 500 was 61.4 percent, on Value Line 29.8 percent. In response to comments by members of the Executive Committee at its meeting on March 22, 1985, Stein Roe sold seven low-yielding stocks with growth potential and bought seven high-yielding stocks in their place. In response to a resolution passed at the annual meeting of the Association on December 29, 1985, Stein Roe sold six stocks doing business in South Africa. Three remain to be sold.

During the past year, the stock market was in two tiers. Large corporations' stocks performed much better than small corporations' stocks. Since the portfolio of the Association is too small for wide diversification among small stocks, its small-stock holdings are in three Stein Roe funds that had total returns of 15.9, 25.0, and 25.4 percent, compared to 21.0 percent for NASDAQ. The Association's return on equities other than (small-stock) funds was 38.2 percent compared to 31.6 percent for Standard & Poor's 500.

Though Mr. McNeill felt that the prospects for the funds were good, he nevertheless proposed reducing their share of the portfolio by one-half. The members of the Finance Committee without expressing strong feelings on the merits of such action decided that the Investment Counsel should have discretion to do so if that was Stein Roe's best judgment.

The Committee made no change in the existing directive specifying that 50 to 75 percent of the portfolio be in equities. It extended from 8 to 10 years the permissible average maturity of fixed-dollar assets excluding the Stein Roe Managed Bonds Fund (which may not be more than 10 percent of the portfolio).

I reported that an officer of the Association had questioned whether the Association needed an investment counsel. He had argued that the Treasurer could invest the equity portion of the portfolio in an S&P 500 fund and the fixed-dollar part in a bond fund. Mr. McNeill pointed out that the purpose of the portfolio as I described it in the minutes of December 12, 1980, was not to maximize returns but to serve "as a reserve held against unforeseen contingencies like the deficit of a quarter of a million dollars incurred in 1970. Accordingly, maintaining the real market value of the portfolio, or at least a substantial part of it, is a major objective." Mr. McNeill said that Stein Roe has been sucessful in attaining that objective. The value of an investment counsel lies in adapting the portfolio to the specific (and sometimes changing) needs of the organization. The discussion continued after the representatives of Stein Roe & Farnham had withdrawn. Without implying dissatisfaction with the performance of Stein Roe, the consensus was that an investigation should be carried out by the Treasurer to verify that the investment counsel is worth the $20,000 a year its services (together with brokers' and custodian's fees) cost. Comparison with the S&P 500 is not enough. A specific plan should be made detailing what the Treasurer would do if he were to manage the portfolio.

Members wanting a list of assets in the investment portfolio can obtain one by writing the Treasurer.

RENDIGS FELS, *Chairman*

Report of the Managing Editor

American Economic Review

Our new decentralized editorial procedures now function quite smoothly and I am pleased to report that they have, in my opinion, had several beneficial effects on our editorial process. The major effect of the new process has been an increased allocation of editorial effort to the appraisal of manuscripts submitted for publication in the *Review* and a commensurate effort devoted to encouraging the expansion and revision of manuscripts considered potentially acceptable. The main cost of this increased editorial effort is a clear and pronounced slowdown in the speed with which we are able to process new submissions.

Editorial Process: As Table 1 indicates, the number of submissions has continued to climb to approximately 1,000 per year. Since we are publishing approximately the same number of papers as we have in the past (see Table 2), this has meant that the chances that a submitted paper will eventually be published have gone down over the past five years.

Table 3, when compared to the data from previous years, provides one indication of the slower rate at which the new editorial process handles manuscripts. In 1984, 77 percent of manuscripts received in a twelve-month period were processed within that period. Table 3 indicates that about one-half of manuscripts received in the period July 1, 1985, through June 30, 1986, had completed processing in this period. Of course, the latter figure is slightly misleading in a period when the number of submissions has been increasing.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1967–86[a]

Year	Submitted	Published	Ratio Published-to-Submitted
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15
1985	952	128	.13
1986	987	123	.125

[a] The submissions reported for every year refer to the last two months of the previous year and the first ten months of the year reported.

TABLE 2—SUMMARY OF CONTENTS, 1985 AND 1986

	1985		1986	
	Number	Pages	Number	Pages
Articles	54	803	55	840
Shorter Papers, including Comments and Replies	74	373	66	344
Dissertations		22		21
Announcements and Notes Section		48		53
Index		10		10
Total		1256		1268

TABLE 3—DISPOSITION OF MANUSCRIPTS, 1985 AND 1986

Manuscripts	March 1– December 31 1985	July 1, 1985– June 30, 1986
Received	726	982
Completed Processing	364	468
Accepted	42	31
Rejected	322	437
Currently in Process	362	514

Table 4 provides the major explanation for why the new editorial process takes longer to reject manuscripts. In Table 4, I provide a breakdown of the distribution of weeks-to-rejection by the number of outside referees used to appraise a paper. It is clear that the major determinant of time-to-rejection is the number of referees asked to appraise a paper. Median weeks to rejection line up as follows:

No Referees = 0–4 weeks;
1 Referee = 15–16 weeks;
2 or 3 Referees = 17–21 weeks.

Under my predecessor, between 30 and 40 percent of papers were handled without the use of an outside referee. As with my predecessor, these are the quick decisions under the new editorial process also. A major change, however, is that I *and* my co-editors have all opted to use a far higher fraction of external referees in the editorial process. Only 18 percent of the papers received in the period March 1, 1985, through June 30, 1986, were handled without consultation with an outside referee.

Table 5 indicates that there has been no change in the speed with which accepted papers are processed *and* published in the *Review*. (June 1986 is the first issue of the *Review* in which papers from the new editorial process were published. September 1986 is the first issue where virtually all the papers published were a result of the new editorial process.) Instead, the delay between receipt and final acceptance of a paper has increased while the delay between acceptance and publication has been reduced. The major cause of the increased delay in the final acceptance of published papers is, in my opinion, the increased revision and expansion of papers that I and my co-editors have requested of the authors of accepted papers. The decreased delay from acceptance to publication reflects the decrease in our backlog of accepted papers.

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION
March 1, 1985–June 30, 1986

Weeks to Rejection	Total Number of Manuscripts	Percent	No Outside Referees	1 Referee	2 Referees	3 or More Referees
0–4	97	.12	86	10	1	0
5–6	74	.09	36	27	11	0
7–8	51	.06	14	26	10	1
9–10	82	.10	7	44	28	3
11–12	61	.07	1	34	21	5
13–14	71	.09	1	37	27	6
15–16	52	.06	0	23	24	5
17–21	129	.16	1	67	54	7
22–26	8	.07	1	23	27	7
27–30	37	.04	3	19	13	2
31–35	48	.06	0	28	18	2
36–52+	65	.08	2	31	22	10
	825[a]	100.	152	369	256	48

[a] The number of rejected manuscripts indicated in this table is larger than the sum of those indicated in Table 3 by a total of 66 manuscripts inherited from the period prior to March 1, 1985.

TABLE 5—AVERAGE PUBLICATION LAGS, BY JOURNAL ISSUE

	Number of Weeks Lag		
Issue	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March 1986	32	47	79
June 1986	41	26	67
September 1986	46	21	67
December 1986	42	25	67

TABLE 6—SUBJECT MATTER DISTRIBUTION OF PUBLISHED MANUSCRIPTS, 1985 AND 1986

	Published	
	1985	1986
General Economics and General Equilibrium Theory	18	9
Microeconomic Theory	12	13
Macroeconomic Theory	15	12
Welfare Theory and Social Choice	8	1
Economic History, History of Thought, Methodology	5	7
Economic Systems	5	1
Economic Growth, Development, Planning, Fluctuations	2	6
Economic Statistics and Quantitative Methods	5	2
Monetary and Financial Theory and Institutions	10	6
Fiscal Policy and Public Finance	5	18
International Economics	5	8
Administration, Business Finance	2	4
Industrial Organization	10	14
Agriculture, Natural Resources	3	1
Manpower, Labor Population	15	13
Welfare Programs, Consumer Economics, Urban and Regional Economics	8	6
Total	128	121

The subject matter distribution of papers published in the *Review* in 1985 and 1986 is contained in Table 6. Having prepared the tabulations for both these years I am surprised they indicate such substantial changes. The classification system used is based on the *Journal of Economic Literature* system and I suspect that the outmoded character of that system induces many arbitrary choices in how articles are classified. It remains my impression that the distribution of published papers reflects fairly accurately the distribution of papers submitted.

One change in the organization of the contents of the *Review* will take place this year. The list of dissertation titles in economics, published in the *Review* since before the advent of the *Journal of Economic Literature*, will henceforth be published in the *JEL*. John Pencavel, Managing Editor of the *JEL*, and I have agreed that the *JEL* is the more appropriate location for this material, especially in view of the energy that must be devoted to classifying these dissertations

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1986 *AER*

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	296	344	$56,922.39	$1,471.14	$58,464
may	28,000	458	496	77,867.17	3,643.73	81,511
June	27,500	290	320	54,399.23	1,620.13	56,019
September	27,500	294	344	59,179.63	1,724.40	60,904
December[a]		388	440	74,800.00	2,200.00	77,000
Annual Misc.[b]						11,783
Total		1,726	1,944	$323,238.42	$10,659.40	$345,681

[a] Estimated.

[b] Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

according to the *JEL* system and the *JEL* staff expertise available for this purpose.

Papers and Proceedings: The ninth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1986. This task was handled by Harvey Rosen (of Princeton University) and Wilma St. John. I am deeply indebted to both of them for the difficult work under extraordinarily tight deadlines that they have so capably performed.

Co-Editors and Board of Editors: I edit the *Review* with the assistance of Robert Haveman, John Riley, and John Taylor. I am deeply indebted to them for the conscientious effort they have expended over the last two years.

The Board of Editors now consists of seventeen members and I am indebted to them also for their efforts. Board members are selected to reflect the highest level of scholarship in the economics profession from the breadth of different fields represented in our submissions. More than fine scholarship is expected of a Board member, however. Board members are also selected because of their conscientiousness, good judgment, and professional reliability. When possible, we like to select Board members from those economists who have been especially helpful in the outside refereeing process.

Two members of the Board completed their terms as of March 31, 1986: George Akerlof and Richard Schmalensee. I am most grateful to them and to the continuing members: Clive Bull, Michael Darby, Jacob Frenkel, Claudia Goldin, Philip Graves, George Johnson, John Kennan, Mervyn King, Meir Kohn, Paul Krugman, Bennett McCallum, Edgar Olsen, Steven Shavell, John Shoven, Susan Woodward, and Leslie Young. I wish also to thank Alvin Roth, a new member of the Board approved for a three-year term beginning on December 31, 1985.

I also wish to thank Wilma St. John, who has continued to perform an outstanding job as production editor, and our editorial assistants Shirley Griesbaum and Sandra Grant, for the fine work they have continued to perform over the past year.

As always, the published version of this report contains the list of referees who have volunteered their services (March 1985 through November 1986). We extend our deepest appreciation for the time and energy they have devoted to the advancement of our science.

A. B. Abel
S. Abizadeh
J. M. Abowd
J. Abraham
K. G. Abraham
I. Adelman
M. Adler
R. Aiyagari
J. Aizenman
G. A. Akerlof
J. Akin
A. Alesi
B. Allen
P. Allen
S. G. Allen
J. Alm
L. J. Alston
J. G. Altonji
E. Alvi
J. E. Anderson
A. K. Ando
R. Andreano
J. Andreoni
J. J. Antel
E. Appelbaum
R. Arnott
K. Arrow
W. B. Arthur
D. A. Aschauer
R. A. Ashley
S. Atkinson
B. K. Atrostic
A. Auerbach
R. Avery
C. Azariadis
D. Backus
K. Bagwell
A. Bahduri
R. Bahl
E. Bailey
M. N. Bailey
Y. Balcer
C. Baldwin
R. Baldwin
I. Ball
E. Baltensperger
W. Barnett
D. Baron
R. J. Barro
J. M. Barron
A. Bartel
T. Bartik
Y. Barzel
L. J. Bassi
R. Batalio
R. Batra
D. S. Batten
C. F. Baum
W. J. Baumol
M. Baxter
C. M. Beach
G. Becker
R. Becker
W. Becker
M. Beckmann
J. Behrman
J. Benhabib
J. P. Benoit
C. Benston
B. Bental
E. Berglas
T. Bergstrom
J. R. Behrman
G. J. Benston
B. S. Bernanke
E. R. Berndt
B. D. Bernheim
S. Berry
H. Bester
D. Betson
T. F. Bewley
J. Bhagwati
S. Bhalla
S. Bhattacharya

M. Bils
J. Bilson
K. Binmore
H. P. Binswanger
R. Bishop
G. Bittlingmayer
S. Black
O. Blanchard
R. M. Blank
D. Blau
B. Blecha
A. Blinder
A. G. Blomqvist
N. Blomquist
D. E. Bloom
M. E. Blume
R. Blundell
R. W. Boadway
D. Bohi
M. Boissiere
L. Boland
P. Bolton
R. Bolton
T. E. Borcherding
K. Border
M. D. Bordo
A. Borges
G. J. Borjas
J. G. Boschen
M. Boskin
R. P. Bosworth
B. L. Boulier
L. Bovenberg
S. Bowles
R. S. Boyer
D. Bradford
R. Braeutigam
W. C. Brainard
B. Branch
J. A. Brander
W. H. Branson
A. Braverman
D. T. Breeden
T. F. Bresnahan
G. Bridgeman
B. Brito
W. Brock
J. Broome
B. W. Brown
C. C. Brown
J. Brown
J. N. Brown

M. Brown
E. Browning
M. J. Browning
D. Brownstone
E. Brubaker
N. Bruce
J. B. Bryant
J. Buchanan
W. H. Buiter
C. Bull
J. Bulow
K. Burdett
A. Burgstaller
R. Burkhauser
G. Burtless
R. Butler
P. D. Cagan
G. C. Cain
G. Calvo
T. Cameron
J. Y. Campbell
R. Cantor
M. B. Canzoneri
A. Caplin
D. Card
M. Carkovic
G. Carliner
C. A. Carlino
J. Carlson
D. W. Carlton
J. Carmichael
L. H. Carmichael
F. Casas
J. Caskey
D. Caves
R. E. Caves
S. G. Cecchetti
K. Chatterjee
J. Chavas
H. Chenery
H. Chernick
A. Chesher
S. N. S. Cheung
R. S. Chirinko
B. Chiswick
G. C. Chow
C. Christ
G. Christainsen
L. R. Christensen
J. Christiansom
L. J. Christiano
L. N. Christofides

C. Cicchetti
S. H. Clain
D. Clark
K. B. Clark
P. K. Clark
C. Clotfelter
A. W. Coats
W. E. Cole
J. L. Coles
D. Collander
J. Conlisk
P. J. Cook
R. W. Cooper
R. D. Cooter
T. Copeland
W. Corden
B. Cornell
P. Courant
D. L. Coursey
R. Craine
P. Cramton
R. W. Crandall
V. P. Crawford
T. Crocker
M. Cropper
A. Cukierman
J. Culbertson
R. Cumby
D. Currie
B. Dahlby
S. Danziger
M. R. Darby
J. Da Vanzo
P. David
S. Davies
R. Day
L. De Alessi
J. B. De Long, Jr.
J. DeMaceda
H. Demsetz
R. Deneckere
E. Denison
M. Denny
A. Denzau
T. F. Dernburg
B. Devaney
D. Diamond
P. Diamond
W. T. Dickens
W. E. Diewert
A. K. Dixit
D. Dollar

J. Donaldson
M. Dooley
R. Dornbusch
M. Dotsey
J. Dow
C. Doyle
A. Drazen
J. Driffill
J. Drisfili
R. A. Driskill
J. A. Dubin
G. J. Duncan
G. M. Duncan
K. B. Dunn
L. F. Dunn
P. H. Dybvig
E. A. Dyl
F. Easterbrook
C. B. Eaton
J. Eaton
B. Eden
S. Edwards
M. Eichenbaum
B. Eichengreen
R. Eisner
P. Elgers
R. Ellis
J. Enelow
M. Engers
R. Engle
D. Epple
N. Ericsson
R. Erikson
M. Eswaran
G. W. Evans
P. D. Evans
S. Fabricant
R. C. Fair
R. Falvey
E. F. Fama
H. S. Farber
R. F. A. Farmer
J. Farrell
R. M. Fearn
R. C. Feenstra
E. Feige
R. Fels
A. Feltenstein
L. Fernandez
C. C. Fethke
G. C. Fethke
A. J. Field

G. Fields
S. Figlewski
R. Filamon
R. K. Filer
R. Findlay
A. T. Finegan
R. Finnie
S. Fischer
F. Fisher
T. Flaherty
R. Flanagan
M. J. Flannery
M. Flavin
C. Flinn
R. P. Flood
R. W. Fogel
J. Formby
R. H. Frank
J. Frankel
H. Frazis
T. Frech III
M. Freeman
R. Freeman
J. A. Frenkel
J. W. Friedman
M. Friedman
K. A. Froot
R. T. Froyen
R. Frydman
V. Fuchs
D. Fudenberg
D. Fullerton
J. Gagnon
F. Gahvari
D. Gale
I. Gale
N. Gallini
H. Galper
P. M. Garber
G. Garvey
D. Gately
J. Geanakopolos
M. Gersovitz
M. Gertler
R. Gertner
J. Geweke
M. R. Gibbons
R. S. Gibbons
R. J. Gilbert
R. F. Gillingham
M. Gisser
A. Glazer

V. Goldberg
A. S. Goldberger
C. Goldin
F. M. Gollop
M. J. Goodfriend
R. Gordon
R. J. Gordon
W. M. Gorman
G. Gorton
P. Gottschalk
J. R. Gould
H. Grabowski
J. Grandmont
K. Grant
P. E. Graves
A. Gray
J. A. Gray
R. Gray
W. H. Greene
C. Greenhalgh
M. Greenhut
B. C. Greenwald
G. Grenier
J. Griffin
J. M. Griffin
Z. Griliches
E. Grinols
R. Gronau
G. Grossman
H. I. Grossman
J. B. Grossman
M. Grossman
S. Grossman
P. Grout
T. Groves
D. Grubb
L. Guasch
R. Guesnerie
A. Gustman
J. Gwartney
D. Haddock
R. W. Hafer
F. Hahn
R. E. Hall
J. C. Haltiwanger, Jr.
J. C. Ham
K. Hamada
D. Hamermesh
B. W. Hamilton
J. D. Hamilton
F. Hammond
P. Hammond

D. Hancock
T. H. Hannan
G. Hanoch
G. D. Hansen
I. P. Hansen
L. Hansen
R. Hansen
E. A. Hanushek
A. Harberger
P. Hare
M. Harris
A. Harrison
D. Harrison
O. Hart
J. Hartigan
P. R. Hartley
D. Hartman
P. T. Hartman
R. Hartman
M. Hashimoto
T. Hatta
J. G. Haubrich
J. C. Hause
J. A. Hausman
T. Havrilesky
J. Hay
F. Hayashi
J. Heckman
A. Heertje
R. Heiner
R. Heinkel
W. P. Heller
M. Hellwig
E. Helpman
W. Hemphill
P. H. Hendershott
D. W. Henderson
J. V. Henderson
W. Hendricks
Z. Hercowitz
J. Hess
D. Hester
E. Hewett
A. L. Hillman
M. Hinick
M. Hirschey
J. Hirshleifer
H. M. Hochman
S. Hoenack
D. Hoffman
W. Hogan
R. M. Hogarth

W. Holahan
R. Holcomb
A. S. Holen
A. S. Holland
S. Hollander
B. Holmstrom
C. A. Holt
D. Holthausen
M. Honig
B. Hool
P. Hooper
K. Hoover
C. Horioka
B. Horrigan
A. J. Hosios
C. Howe
P. Howitt
W. Hoyt
C. Hsiad
R. G. Hubbard
P. Hughes
J. Huizinga
R. M. Hutchens
C. Ichniowski
Y. Ioannides
J. Ingersoll
R. Inman
R. M. Isaac
P. Isard
T. Itagaki
T. Ito
A. B. Jaffe
G. Jakubson
M. Jensen
R. Jensen
W. Joerding
A. John
D. G. Johnson
G. Johnson
P. A Johnson
R. Johnson
T. Johnson
W. R. Johnson
R. A Jones
R. W. Jones
S. Jones
P. L. Joskow
B. Jovanic
K. Judd
R. Just
H. Juster
J. Kagel

C. Kahn
L. Kahn
D. Kahneman
J. Kalt
M. Kamien
E. Kane
J. Karekan
S. Karlson
E. Karni
E. Katz
L. F. Katz
M. Katz
D. Katzner
J. Kau
B. Kaufman
H. Kaufman
S. Kealhofer
T. Keeler
P. Kehoe
M. Kemp
J. Kendrick
J. Kenen
J. Kennan
J. Kesselman
N. M. Kiefer
R. Kihlstrom
M. Killingsworth
Y. Kim
M. King
S. King
P. Kino
I. Kirzner
N. Kiyotaki
A. W. Kleidon
B. Klein
T. J. Kniesner
B. Kobayashi
L. Kochin
R. Koenker
R. Koller
A. Koo
R. W. Kopcke
R. Kormendi
L. Kotlikoff
P. Kouri
M. Kraus
G. Krause-Junk
M. Krauss
I. Kravis
K. Krishna
C. Krouse
P. Krugman
P. Kumas
M. Kurz
F. Kyoland
A. Kyle
P. Kyle
J. J. Laffont
K. Lahiri
D. F. W. Laidler
R. J. LaLonde
R. Lambert
V. Lambson
K. Lang
W. W. Lang
L. J. Lau
R. Layard
E. Lazear
E. E. Leamer
R. D. Lee
T. Lee
W. Lee
N. H. Leff
K. Leffler
R. H. Leftwich
B. N. Lehman
D. Lehman
D. Leigh
J. P. Leigh
L. S. Leighton
A. Leijonhufvud
J. Leland
J. S. Leonard
M. D. Levi
D. Levin
D. Levine
F. Levy
A. Lewis
H. Lewis
T. Lewis
L. Li
D. Lilien
L. Lillard
R. C. Lind
C. Lindsay
A. Link
P. Linneman
S. A. Lippman
R. Lipsey
R. Litterman
R. Litzenberger
J. Loeys
R. Lombra
S. Long
W. Long
G. Loury
M. C. Lovell
R. E. B. Lucas
R. F. Lucas
M. Lundahl
S. Lundberg
P. McAfee
M. McBride
T. McBride
J. J. McCall
B. T. McCallum
C. McCann
D. McCloskey
C. E. McClure
T. McCool
R. McCulloch
J. B. McDonald
J. M. McDowell
K. M. McElroy
M. B. McElroy
D. McFadden
J. McIntosh
R. McKelvey
G. McKenzie
R. McKinnon
A. M. McLennan
J. McMillan
M. McMillan
N. McMullen
L. Maccini
M. Machina
R. MacKay
D. MacRae
T. E. MaCurdy
G. S. Maddala
A. Maddison
W. Magat
S. Magee
G. Mailath
J. H. Makin
N. G. Mankiw
M. Mann
R. Manning
C. Manski
R. Manuelli
A. Marcus
D. Margaritis
S. Marglin
R. A. Margo
J. R. Markussen
T. A. Marsh
J. Marshall
R. Marston
H. Marvel
S. E. Masten
F. Mathewson
D. J. Mathieson
R. C. O. Matthews
S. Matthews
S. Matusz
T. Mayer
J. L. Medoff
R. A. Meese
A. Melino
A. H. Meltzer
J. Melvin
M. T. Melvin
P. L. Menchik
R. Merton
M. Meruer
L. Meyer
R. Michael
P. Mieszkowski
J. Migue
H. Milde
P. R. Milgrom
M. Miller
P. Miller
D. Mills
J. Mince
J. Mincer
T. Mirer
L. Mirman
J. Miron
E. Mishan
F. S. Mishkin
D. J. B. Mitchell
O. Mitchell
D. Modest
F. Modigliani
R. Moffitt
R. Mohan
J. Mokyr
M. Montgomery
J. Montias
J. Moore
W. J. Moore
P. J. Morgan
S. A. Morley
R. Morris
C. J. Morrison
D. Mortenson
O. T. Mortenson

L. N. Moses
T. Mroz
J. N. Muellbauer
D. Mueller
P. Mullineaux
C. Mulvey
K. Murphy
K. J. Murphy
M. Murray
R. Musgrave
R. Muth
S. Myers
R. Myerson
B. Nalebuff
P. Neary
R. Neary
S. N. Nefici
C. R. Nelson
J. Nelson
R. Nelson
R. A. Nelson
L. Neuberg
D. Newbery
W. Newey
S. Nickell
D. Nichols
L. Nichols
D. R. Nickerson
M. Nishmizu
R. Noll
W. Nordhaus
V. Norman
R. Norsworthy
W. Novshek
W. H. Oakland
W. Oates
R. Oaxaca
A. O'Brien
M. Obstfeld
G. P. O'Driscoll
K. Ohno
H. Ohta
W. Oi
S. Oliner
E. Olsen
R. Olsen
M. Olson, Jr.
Y. Ono
J. Ordover
D. Osborne
M. Osborne
S. Oster
J. Ostroy
A. J. Oswald
S. Ozler
A. R. Pagan
M. Paglin
A. Pakes
R. Palmquist
J. Panzar
O. Papell
R. W. Parks
D. O. Parsons
W. Parys
M. Pascoa
P. Pashigan
J. D. Paulus
M. Pauly
S. Payne
J. Pechman
M. Peck
S. Peltzman
J. Pelzman
J. H. Pencavel
E. Pentecost
J. M. Perloff
G. L. Perry
M. Perry
B. C. Petersen
L. Philps
P. J. Pieper
D. A. Pierce
J. I. Pierce
J. Pincus
R. Pindyck
J. Pippenger
R. Piron
C. A. Pissarides
M. Plant
G. Plesko
R. Plotnick
C. Plott
I. P'ng
S. W. Polachek
H. M. Polemarchakis
A. J. Policano
M. Polinsky
R. Pollak
W. Poole
R. Pope
R. Porter
P. Portney
M. Post
A. Postlewaite
J. M. Poterba
E. C. Prescott
A. Protopapadakis
S. Prowse
T. Pugel
D. H. Pyle
R. E. Quandt
R. Radner
D. Rae
J. Raisian
K. Ramaswamy
V. Ramey
A. J. Randall
G. Ranis
S. I. Ranney
M. R. Ransom
R. H. Rasche
A. Raskovich
E. Rasmusen
R. Raucher
R. Rauner
D. Ravenscraft
A. Raviv
S. Rea
P. Reagan
A. E. Rees
D. Rehm
M. Reich
C. E. Reid
C. Reimers
J. Reinganum
U. Reinhardt
P. C. Reiss
F. Remington
R. Reynolds
D. Richardson
W. C. Riddell
W. Riker
I. Rima
M. Riordan
J. Ritzen
F. Rivera-Batiz
R. Rob
J. Roback
D. J. Roberts
J. M. Roberts
M. C. Roberts
C. Robinson
A. Robson
K. Rock
C. Rodrigues
D. Rodrik
C. Rogers
R. Rogerson
W. Rogerson
K. S. Rogoff
V. Roley
R. Roll
A. Rolnick
C. Romer
D. Romer
T. Romer
S. Rose-Ackerman
H. Rosen
K. T. Rosen
S. Rosen
M. Rosenzweig
S. Ross
T. Ross
R. Rossana
J. Rotemberg
A. Roth
M. Rothschild
R. Rothschild
F. Rozwadowski
R. S. Ruback
D. L. Rubinfeld
A. Rubinstein
J. Rudin
R. Ruffin
M. Rush
L. B. Russell
T. Russell
G. Russo
J. Rust
V. W. Ruttan
P. A. Ruud
H. E. Ryder, Jr.
J. Sachs
R. K. Sah
D. Saks
M. Salami
G. Salamon
S. Salant
G. Saloner
S. Salop
P. Samuelson
W. Samuelson
W. Sander
A. Sandmo
A. M. Santomero
D. Sappington
T. Sargent
R. Sato

M. Satterthwaite
P. Saunders
W. Saupe
J. Scadding
D. Scharfstein
D. Scheffman
T. Schelling
F. M. Scherer
J. Schiff
D. Schagenhauf
M. Schlesinger
R. Schmalensee
P. Schmidt
A. Schmitz
J. Schmitz
T. Schneeweis
M. S. Scholes
A. Schotter
S. Schwab
A. J. Schwartz
M. Schwartz
G. W. Schwert
S. Scotchmer
J. Seade
J. J. Seater
G. Sedlacek
U. Segal
L. Seidman
L. Selwyn
R. Shakotko
D. R. Shaller
C. Shapiro
W. Sharkey
S. Sharma
S. Shavell
S. M. Sheffrin
L. Shepard
W. Shepherd
R. Sherman
E. Sheshinski
R. Shiller
J. Shoven
W. Shughart
R. C. Sickles
T. Sicular
D. Siegel
J. Siegel
E. Silberberg
C. Simon
L. Simon
C. A. Sims
N. Singh
K. J. Singleton
A. Skinner
J. Skinner
F. Sloan
K. Small
T. Smeeding
A. Smiley
M. Smirlock
B. D. Smith
G. Smith
J. P. Smith
L. Smith
R. S. Smith
V. Smith
V. K. Smith
V. L. Smith
E. Smolensky
D. Snower
E. Solomon
G. R. Solon
R. Solow
H. Sonnenschein
R. Spady
B. Spencer
D. F. Spencer
M. Spiegel
R. G. Spigelman
D. Spulber
T. N. Srinivasan
R. Staaf
J. Staddon
F. P. Stafford
R. Staiger
O. Stark
R. Startz
H. Stein
J. Stein
P. E. Stephan
N. Stern
J. Stewart
M. Stewart
M. D. Stewart
G. Stigler
J. E. Stiglitz
J. Stock
T. Stocker
J. Stockfish
A. Stockman
T. M. Stoker
N. I. Stokey
C. Stone
R. Strauss
C. Stuart
A. Sullivan
D. Sullivan
A. Summers
L. Summers
R. Summers
J. Sutton
J. Svejnar
L. E. O. Svensson
L. G. Svensson
J. L. Sweeney
R. E. Sylla
M. Syrquin
G. Tabellini
K. Taira
W. Takacs
P. Tandon
E. Tanner
D. Tarr
J. Tatom
P. Taubman
J. Taylor
L. Taylor
D. Teece
L. Telser
R. Thaler
D. Thomas
H. L. Thompson
L. C. Thurow
N. Tideman
J. Tirole
R. Tisinai
S. Titman
J. Tobin
E. Todaro
L. S. Topel
R. Topel
E. Tower
R. M. Townsend
J. S. Tracy
R. Tresch
R. Triest
J. E. Triplett
B. Trueman
T. J. Trussell
G. Tullock
J. Turner
S. J. Turnovsky
L. Tyson
A. Ulph
D. Usher
J. Van der Gaag
F. Van Winden
H. Varian
T. Venables
R. Verreccia
D. Vines
H. D. Vinod
W. K. Viscusi
T. Vishwanath
X. Vives
P. Voos
P. A. Wachtel
M. Wachter
M. Waldman
D. G. Waldo
T. Wales
I. Walker
N. Wallace
K. F. Wallis
C. E. Walsh
H. Wan
G. H. Wang
M. Watson
L. Waverman
P. Weil
B. Weingast
B. Weisbrod
A. Weiss
L. Weiss
Y. Weiss
M. Weitzman
S. Wellisz
E. G. West
K. West
F. Westfield
L. Westphal
J. Whalley
M. D. Whinston
L. H. White
M. J. White
C. H. Whiteman
E. Wicker
J. A. Wilcox
J. A. Wilde
L. Wilde
J. T. Williams
J. Williamson
O. Williamson
R. Willig
C. A. Wilson
J. Wilson
R. B. Wilson
G. Winston

R. Winter
S. G. Winter
D. Wise
B. Wolfe
E. Wolff
K. Wolpin
R. Wonnacott
W. T. Woo
D. Wood
S. Woodbury
M. Woodford
S. Woodward
J. M. Wrase
B. Wright
G. Wright
M. Wright
M. Yaari
J. Yellen
J. Yinger
P. A. Yotopoulos
D. Young
L. Young
S. D. Younger
G. A. Zarkin
V. Zarnowitz
R. Zeckhauser
S. P. Zeldes
A. Zellner
R. Zerbe
A. Zimbalist
M. Zimmerman
G. R. Zodrow

ORLEY ASHENFELTER, *Managing Editor*

Report of the Managing Editor
Journal of Economic Literature

At the beginning of 1986, the editorship of the *Journal* changed hands. As indicated in the Editor's Note in the March 1986 issue, this change in Managing Editor will not mean a change in the *Journal*'s basic purpose which is to provide economists with a comprehensive guide to economics research and publications. This guide takes the form of critical survey and review articles, book reviews, and an annotated listing of new books, a compendium of tables of contents of current periodicals, and selected abstracts of articles. The classified indexes of articles appearing in the quarterly issues of the *Journal* are assembled in an annual *Index of Economic Articles*. This *Index* also includes papers that have appeared in collected volumes such as conference proceedings and *Festschriften*. In general, the *Journal* aims to help members of the American Economic Association maintain a broad knowledge of economics in the face of powerful pressures toward specialization.

The division of the *Journal*'s pages among the major departments since 1980 is given in Table 1 from which it is evident that it has been primarily the growth in the listing of the contents of current periodicals that has accounted for the expansion in the *Journal* during the last six years. This growth in the *Journal*'s bibliographic work reflects the increase in the number and size of economics periodicals throughout the world. As noted in previous Managing Editor's Reports, this growth presents a problem for the *Journal*: how can we maintain the usefulness of these bibliographic services without substantially increasing the *Journal*'s size and expense? In the last few years, surveys have been undertaken to review the periodicals listed in the *Journal* and this has helped to contain the growth in the periodicals department. Nevertheless, further steps may be called for if current trends continue.

The bibliographic departments (i.e., the Annotated Listing of New Books, the Contents Periodicals, and Selected Abstracts of Articles) are run from the *Journal*'s Pittsburgh office under the direction of Drucilla Ekwurzel. Linda Scott serves as Assistant Editor and Professor Asatoshi Maeshiro of the University of Pittsburgh acts as Editorial Consultant with general responsibility for the classification of articles on which the *Journal*'s subject index depends. In 1986, the *Journal* provided annotations of over 1,200 new books, listings of more than 1,200 issues of various economics journals, and abstracts of over 2,400 articles. Also the 1983 *Index of Economic Articles* was published with the 1981 *Index* prepared for publication. The *Index* for 1982 will follow shortly. The subject and author indexes of journal articles are available through computer access to the *Economic Literature Index* (*ELI*) of the DIALOG Information Retrieval Service. An article by Drucilla Ekwurzel and Bernard Saffran in the December 1985 issue of the *Journal* provides information about this service.

The Articles and Communications and Book Review departments of the *Journal* are managed from the *Journal*'s Stanford University office. During 1986, the *Journal* published 8 major articles, 3 communications, and 166 book reviews. Alex Field supervises the Book Review department and Moses Abramovitz shares with me the editorial duties associated with reviewing the manuscripts. We have been assisted by an able and conscientious Board of Editors and by almost 100 referees and reviewers. Seven members of the Board have completed their terms as of the end of 1986, namely, Carolyn Shaw Bell, Robert Eisner, Duncan Foley, Victor Fuchs, Jack Hirshleifer, David Laidler, and Roger Noll. I am most grateful for their help. I am very pleased that Robert Eisner, Duncan Foley, David Laidler, and Roger Noll have agreed to serve another term on the Board. I shall propose that Mark Killingsworth join the Board next year.

TABLE 1—*JEL* PAGES BY DEPARTMENT: 1980–86

	Articles and Communications	Book Reviews	New Book Annotations	Current Periodicals	General Index	Total
1980	366	294	276	1,072	26	2,034
1981	342	286	270	1,059	23	1,980
1982	331	251	300	1,069	23	1,974
1983	305	239	281	1,086	38	1,949
1984	354	225	314	1,193	37	2,124
1985	364	237	299	1,306	38	2,244
1986	367	250	308	1,344	38	2,307
△1980/1–1985/6[a]	+3.2	−16.0	+11.2	+24.4	+55.1	+13.4

[a] Denotes the percentage increase from 1980/1 to 1985/6.

I thank all members of our Board for their very real contributions to the *Journal*. I am also indebted to our referees who have helped determine and develop the *Journal*'s articles.

The punctual and successful production of the *Journal* in 1986 has depended on the efforts of many other people. At Pittsburgh, the full-time members of the staff include Pat Andrews and Beth Thornton. The Stanford office relies on the efforts of Anita Makler, Ann Vollmer, and Toni Haskell. I am grateful for their devoted and effective work.

JOHN PENCAVEL, *Managing Editor*

Report of the Editor

Journal of Economic Perspectives

I am pleased to report that the new journal, established by the Executive Committee of the American Economic Association at its March 1986 meeting, seems well on its way to a successful beginning. We have chosen the name *Economic Perspectives* to capture the journal's twin missions of providing perspective on current economic research, and explaining how economics provides perspective on questions of general interest.

The first task after the authorization of the new journal was the appointment of a co-editor. I was fortunate to be able to persuade Carl Shapiro, of the Woodrow Wilson School, Princeton, to be the co-editor. Carl, besides being an excellent economist, has had extensive editorial experience; he had been associate editor of the *Rand Journal* and the *Quarterly Journal of Economics*, and had just been asked to become co-editor of the *Rand Journal*. He received rave reviews from the editors with whom he had worked on these journals, which, on the basis of our work together over the past few months, were well deserved.

Our second task was to establish an editorial office. We were again most fortunate in being able to hire Timothy Taylor as our managing editor. After two years of graduate training in economics at Stanford, he had worked for several years writing editorials and feature articles on economics for the *San Jose Mercury News*. We also were fortunate in finding an editorial associate, Caroline Moseley, who could not only efficiently manage the day-to-day affairs of the office, but also do copyediting and perform other editorial functions.

Office space in Princeton is expensive and hard to find, and, until the completion of a new building which the university is constructing for the Economics Department, space within the university is hard to come by. We were therefore most pleased by the efforts of Donald E. Stokes, Dean of the Woodrow Wilson School of Public and International Affairs, who secured us offices in the Woodrow Wilson School. Being able to avail ourselves of the facilities of the School has also greatly simplified the tasks of organizing our offices.

Our third task was to appoint an editorial board. We sought individuals who had demonstrated an interest in and ability to communicate ideas to a wide circle of economists. We sought to achieve a balanced coverage of subdisciplines within economics, and a balanced representation by geographical and other characteristics. We were extremely pleased by the enthusiasm with which the proposed new journal was received by those we approached. Our editorial board is as follows: Henry J. Aaron, The Brookings Institution; Stanley Fischer, Massachusetts Institute of Technology; Paul R. Krugman, Massachusetts Institute of Technology; Edward P. Lazear, Graduate School of Business, University of Chicago; Mark J. Machina, University of California-San Diego; Charles F. Manski, University of Wisconsin; Donald N. McCloskey, University of Iowa; Bernard Saffran, Swarthmore College; Steven C. Salop, Georgetown University Law Center; Lawrence H. Summers, Harvard University; Hal R. Varian, University of Michigan; Janet L. Yellen, School of Business, University of California-Berkeley.

Our next task was the solicitation of papers. The interactive process that we had originally envisaged worked extremely well. The editor and co-editor had extensive discussions with each other and with each of the associate editors to establish a long list of papers and symposia, and of potential authors for each. These were narrowed to a short list of those to be approached immediately, and a somewhat longer list of those to be approached within the near future. These were followed up by discussions with the authors, explaining both the purposes of the new journal as well as the article that we were soliciting. In almost all cases, the authors were not only willing to write the article, but were enthusiastic about the project. Our associate editors again did an excellent job in providing comments to the authors, and the interactions between our editorial office and the associate editors went extremely smoothly. (Our first issue will include a special symposium on tax reform, with articles by Charles E. McLure, Jr. and George R. Zodrow, James Buchanan, Richard A. Musgrave, Alan Auerbach, Jerry Hausman and James Poterba, and Paul Courant and Daniel Rubinfeld; a symposium on option pricing, with articles by Varian and by Rubinstein; and articles by Gavin Wright, Mark Machina, and Lawrence Summers. It will also include several special features that our associate editor, Saffran, has organized.)

We are now in the midst of our final two tasks. Almost all of the articles solicited for the first issue arrived within a week or two of schedule. These have been returned to the authors for revisions. All of the authors had made a real attempt to communicate their ideas to a wide audience, and almost all of them were successful. All of the authors have also been extremely responsive to our suggestions for revision.

Our final task is the production of the new journal. This entails making a number of decisions concerning the design specifications of the journal and its cover, and soliciting and evaluating bids from printers. This task should be completed by early January.

Although we do not yet have a backlog of completed articles, if we receive the same kind of cooperation from those authors who have agreed to write papers over the next few months that we received from those who agreed to contribute papers to our first issue, we should be able to maintain our production schedule. We plan to publish two issues in 1987, and to issue the journal quarterly thereafter. The first issue is scheduled to come out in midsummer.

Report of the Director

Job Openings for Economists

The total number of new jobs listed this year was 1,561. In 1985, the number was 1,592 and in 1984, 1,713. Both academic and nonacademic listings decreased slightly from 1985 to 1986. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type (academic and nonacademic) for each issue of *JOE* in 1986.

Table 2 shows the number of employers by category (four-year colleges, universities with graduate programs, federal government, etc.) for each of the 1986 issues. Academic

TABLE 1—JOB LISTINGS FOR 1986

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	59	106	46	76
April	48	78	45	73
June	27	45	26	42
August	41	98	38	90
October	155	399	131	356
November	132	285	132	285
December	201	465	95	212
Subtotal	663	1,476	513	1,134
Nonacademic				
February	13	60	10	49
April	17	57	16	52
June	13	40	11	30
August	13	52	12	47
October	29	107	27	101
November	19	82	19	82
December	53	172	30	66
Subtotal	157	570	125	427
TOTAL	820	2,046	638	1,561

TABLE 2–NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1986

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Govern-ment	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	23	36	2	1	1	2	5	2	72
April	12	36	1	3	–	3	7	3	65
June	9	18	2	1	2	–	7	1	40
August	16	25	2	–	1	2	6	2	54
October	48	107	9	1	4	1	10	4	184
November	48	84	6	2	2	–	8	1	151
December	79	122	15	3	13	2	17	3	254
TOTAL	235	428	37	11	23	10	60	16	820

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1986

Fields[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	65	45	24	66	183	143	229	755
Growth and Development (100)	11	11	9	19	41	26	51	168
Econometrics and Statistics (200)	17	20	15	23	70	54	102	301
Monetary and Fiscal (300)	29	25	15	20	92	56	124	361
International Economics (400)	12	12	6	18	68	43	91	250
Business Administration, Finance, Marketing and Accounting (500)	13	8	9	16	41	26	49	162
Industrial Organization (600)	15	15	13	16	50	34	78	221
Agriculture and Natural Resources (700)	9	10	7	8	17	22	22	95
Labor (800)	10	9	8	14	48	28	58	175
Welfare and Urban (900)	6	10	4	5	33	37	48	143
Related Disciplines (A00)	1	5	1	1	6	5	12	31
Administrative Positions (B00)	5	4	2	2	7	9	17	46
TOTAL	193	174	11	208	656	483	881	2,708

institutions continue to be the major source of jobs, about 80 percent of the total number of employers listing vacancies. The pattern this year is basically the same as it has been for several years.

Table 3 shows the number of citations by field of specialization. General economics (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200). This continues the pattern that has prevailed for several years.

The Association is fortunate to have Violet Sikes. She is responsible for everything to do with the publication and distribution of *JOE*. I am indebted to her for the excellent work she continues to do.

C. ELTON HINSHAW, *Director*

Report of the Committee on Economic Education

A two-year-long collaborative effort by this Committee and the Joint Council on Economic Education resulted in recent action by the College Board approving the establishment of an Advanced Placement Program in Economics. This approval, coming in December 1986, means that the program will get underway in the 1988–89 academic year, with the first test administration in the spring of 1989. Advanced Placement (AP) Programs in a variety of other high school subjects have existed for many years, facilitated and development of college-level courses in the high schools, and enabled college preparatory students who do well in these courses, as evidenced by the AP examinations, to obtain college credit for their high school work. We have pushed for development of a similar program in economics in order to give more prominence to a subject whose teaching is now mandated in more than half the states, to enhance the quality of teaching in high school economics, and to encourage the development of curriculum materials and testing instruments that will improve the quality of the economics courses offered in the high schools.

Because some economists have expressed reservations about this development, a meeting of department chairs from some of the major universities was convened at the December American Economic Association meetings to review the matter. Those attending the meeting, hosted by the New Orleans Branch of the Atlanta Federal Reserve Bank, were briefed on the AP Program by economist Stephen Buckles who chairs the committee for the AP Program in economics and by Harlan Hansen from the College Board who oversees the AP Programs and who was instrumental in gaining approval for the economics program. The general reaction among those attending was highly favorable. Of course, economists will be fully involved in the development of the AP Program, and they will play a major role in formulating the AP examinations. Ultimate responsibility for granting college-level credit for AP work rests with individual colleges and universities and within them usually with economics departments. Additional meetings of the kind held in New Orleans are planned for subsequent AEA meetings so as to keep the profession informed about the program and its evolution.

The AP Committee is now at work developing the syllabus and preliminary versions of the AP test. The Alfred P. Sloan Foundation has already provided a sizeable grant to assist in implementing the AP Program. The College Board will also be making a substantial investment in test development, teacher training sessions, and publicizing the program.

Some concern surfaced at the Committee's meeting about the apparently growing fraction of foreign students in economics graduate programs and the problems that arise when these students receive appointments as teaching assistants. While they are usually well-equipped technically, their language difficulties and lack of cultural orientation to the United States and to American students frequently impedes their teaching effectiveness. The Committee will make a preliminary exploration of this problem, perhaps in conjunction with the AEA-sponsored Economics Institute located at the University of Colorado, Boulder. The Institute regularly offers intensive and highly effective language study combined with special prepatory courses in economics for foreign students planning to enter U.S. graduate programs in economics, agricultural economics, and business. One possible solution would be to encourage foreign students entering Ph.D. programs to enroll in the Institute during the summer prior to beginning graduate study here, particularly if their language skills are deficient. The record of the Institute in improving TOEFL scores, for example, is most impressive.

There was also some discussion about what might be done to improve the assessment of student learning in the economics major. Questions about what students are learning in college are receiving increased attention in various reports on the state of higher education that have surfaced in the past several years. Because the Joint Council on Economic Education is in the process of revising its various tests, this is an ideal time to think about the possibility of broadening the scope of its testing instruments. One important continuing item on the Committee's agenda is the task of trying to secure funding for updating the Teacher Training Program; this has been discussed more fully in the past several Committee reports.

It should be reported that with the support of the J. Howard Pew Freedom Trust, information is being gathered on the level of economic understanding possessed by a national sample of high school seniors. The Test of Economic Literacy is being used to generate baseline data in this project. The purpose of the project which is directed by William J. Baumol and Robert J. Highsmith is to provide new data so as to both stimulate and facilitate research in economic education. In due time the data will be made available to researchers for further analysis.

Finally, the Committee received a report on the health and direction of the *Journal of Economic Education* (*JEE*) from its editor, Kalman Goldberg. Special efforts are being made to increase the circulation of the *JEE* through an extensive promotional mailing this spring. The quantity of manuscripts submitted has continued to increase, resulting in some backlog of accepted papers. In addition, the Spring 1987 issue will feature the conference proceedings of the recent MIT Conference on The Scope of Economics which was supported by the Calvin K. Kazanjian Economics Foundation, Inc.; Robert Solow chaired the planning committee for the conference. Sometime early in 1988 the *JEE* will feature a symposium on economics textbooks. The Committee is pleased with the broader orientation the *JEE* is taking and invites members of the profession to not only subscribe but also submit papers to it.

Policy and Advisory Board of The Economics Institute

During 1986, the Economics Institute has continued to carry out its mission as the American Economic Association's institution for preparatory training for foreign students coming to the United States to earn advanced degrees in economics, agricultural economics, business, and administration.

During recent years, almost all the Institute's revenues have come from students' tuition and fee payments. In 1986 there was a continuation of slow growth and foreign exchange shortages in Third World countries, from which nearly all the Institute's students come. Nevertheless, the Institute again trained more than 400 students prior to entry into U.S. graduate programs. These students came from more than 50 countries and enrolled in more than 75 U.S. universities upon completion of their studies at the Economics Institute. By far the most serious problem facing the Institute is its drastic shortage of scholarship funds to help finance students from countries which cannot fully finance their students. Lack of funds prevents many qualified students, especially from South Asia and Africa, from enrolling in the Institute and proceeding to successful graduate study in U.S. universities.

Foreign students have become extremely important in U.S. graduate programs in economics. About one-third of all advanced degrees awarded by U.S. universities in economics are received by foreign students. Similar percentages appear to be accurate for both M.A. and Ph.D. degrees, and appear to span the geographical and quality spectra of U.S. universities. U.S. graduate education has contributed materially to the quality of economic research, instruction and government policymaking in many Third World countries. In addition, many U.S. economics departments depend on foreign students for adequate enrollments in graduate programs and for instructors in undergraduate courses.

The impact of the Institute is to improve the cost effectiveness of U.S. graduate training for foreign scholars. For nearly 30 years, it has addressed this task through highly professional, innovative, and carefully tailored programs of instruction in English, economic theory, mathematics and statistics, accounting, finance and management, and more recently in computer competency and applications. For large numbers of students, quality university placements have also been facilitated. Both the time and cost required to obtain advanced degrees in U.S. universities are minimized by pre-training and associated orientation to U.S. campus and community life at the Economics Institute. Training at the Institute also enables foreign students to function more effectively as teaching assistants and instructors at their universities.

The Institute is expanding collaborative arrangements with graduate degree programs designed to help additional individual foreign scholars, and organizations with staff development programs in other countries, to more effectively access graduate degree programs in the United States and to complete programs of study with improved academic and cost records. Interested programs and organizations are invited to contact the Director, 1030 13th Street, Boulder, CO 80302 (telephone 303 492–8419; Telex 450385 ECONINST).

WYN F. OWEN, *Director*

EDWIN S. MILLS, *Chairman*

Report of the Representative to the American Association for the Advancement of Science

There has not been a formal report by the AEA Representative to the American Association for the Advancement of Science (AAAS) for several years, so I believe it is appropriate to reacquaint the membership with the operation of the AAAS, in addition to reporting on the activities during the past year that involved economists. The AAAS is a federation of scientific organizations, as well as an association of over 135,000 individual members. Its objectives are "to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress."

Since its founding in 1848, economists have generally played a limited role in the AAAS, though three economists (Carroll Wright, Wesley Mitchell, and Kenneth Boulding) have ascended to the organization's presidency. The discipline of economics is grouped with other social sciences in Section K, one of the twenty-one directorates of the AAAS. The approximately 360 economists who belong to Section K represent just under 20 percent of the section's membership.

Participation in the AAAS offers economists a chance both to learn from and to contribute to other disciplines. However, many AEA members have noted that the AAAS is much more multidisciplinary than interdisciplinary, and, given the extent of scientific specialization, the opportunities for interaction may thus be quite limited. This has been evidenced by the major AAAS publication, *Science*, which is dominated by the physical and biological sciences, and which many economists have found inaccessible as readers or as authors of papers.

A concerted effort has been underway by leaders of both organizations over the last few years to improve this situation. During 1986, Robert Solow was appointed to the Editorial Board of *Science*. This led to a record number of papers by economists, many of which were commissioned, appearing in the publication. Authors included Paul Samuelson, Charles Plott, Richard Cooper, Vernon Smith, Martin Baily, Rudiger Dornbusch, Stanley Fisher, Franco Modigliani, and Elizabeth Bailey. Topics were not confined to those of interdisciplinary interest, such as experimentation and the economics-technology interface, but also included "Third World Debt" and "Deregulation."

The 1986 annual meeting of the AAAS was held in Philadelphia in May. No session at the meeting was explicitly or exclusively devoted to economics (though there will be two designated economics sessions at the 1987 meeting). Nevertheless, economists participated in several ways. Kenneth Boulding presented a Plenary Lecture, and Lester Thurow participated in a session entitled, "The Frontiers of the Social Sciences: Integration of the Economic and Other Social Sciences." There were several sessions of related interest to the AEA membership, such as those on agricultural policy, economic aspects of demographic trends, and natural resources. The meeting also included sessions of general interest, such as "Scientific Freedom and Responsibility" and "Science: Education and Public Understanding."

The AAAS co-sponsored two sessions at the Allied Social Sciences Meetings in New Orleans. One was a joint AEA/AAAS session entitled, "Technology and Employment." Another, held jointly with the Association of Environmental and Resource Economists, was entitled "Managing Environmental Risk."

I began the duties of the AEA representative to the AAAS for a three-year term in February 1986. On behalf of the AEA, let me take this opportunity to thank my predecessor, Roger Bolton of Williams College, for his service over the course of two terms between 1980 and 1986.

Further information about the American Association for the Advancement of Science is available from Marge White, AAAS, 1333 H Street, N.W., Washington, D.C. 20005.

ADAM ROSE, *Representative*

Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research conducts analyses on a large variety of economic issues; publishes books, working papers, and two periodicals; sponsors conferences; and holds workshops and seminars as part of an annual summer institute. Approximately 280 economists at universities across the United States contribute to NBER's working paper series, and many other economists, here and abroad, attend conferences and the summer institute.

Programs. NBER's research is organized into eight programs (directors in parentheses): Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity and Technical Change (Zvi Griliches). Program meetings are generally held twice during the academic year and once, for a longer period, during the summer institute. About 316 people attended summer institute meetings in Cambridge in 1986.

Projects. The NBER also sponsors large-scale projects which bring together researchers from several of these programs. One major project centers on the government budget and the private economy. It includes the following subprojects (directors in parentheses): the impact of taxation on such behavior as charitable contributions (Charles Clotfelter); measuring and analyzing the role of state and local government in the economy (Harvey Rosen); studies of the compensation of public sector employees (David Wise); the impact of public sector unionization (Richard Freeman); and an analysis of government debt and deficits and their impact on the private sector (David Bradford and Benjamin Friedman).

In addition, NBER is currently sponsoring several smaller projects. William Branson and J. David Richardson are directing a project on international economic policy. Research on trade relations and trade policy is directed by Robert Baldwin. Richard Marston leads a group studying the effects of misaligned exchange rates.

Stanley Fischer is coordinating a project on macroeconomic policy. Alvin Klevorick is organizing a study of strategic trade. Jeffrey Sachs is directing a project on developing country debt. Richard Freeman heads a study of international migration. Alan Auerbach is leading an examination of mergers and acquisitions. Martin Feldstein is directing a study of the effects of taxation on capital formation. And David Wise heads a project on the economics of aging.

Conferences. In 1986, NBER sponsored conferences (organizers in parentheses) in the United States and abroad on the following topics: Political Economy (Robert Baldwin and Robert Keohane), Effects of Taxation on Capital Formation (Martin Feldstein), Macroeconomics (Stanley Fischer), International Economic Problems in the U.S. and Japan (Koichi Hamada and Richard Marston), Empirical Methods for International Trade (Robert Feenstra), Keynesian and Classical Economics After 50 Years (Bennett McCallum), Taxes and Capital Formation (Martin Feldstein), Europe–U.S. Trade Relations (Robert Baldwin and André Sapir), International Seminar on on Macroeconomics (Robert Gordon and Georges de Menil), International Asset Pricing (Bernard Dumas), Economic Issues in the U.S. and Japan (Geoffrey Carliner), Public Sector Unionism (Richard Freeman), Mergers and Acquisitions (Alan Auerbach), Government Expenditure Programs (Jerry Hausman and James Poterba), and State and Local Government Finance (Harvey Rosen).

During 1986 the NBER published six books: *Financing Corporate Capital Formation* (Benjamin Friedman, ed.); *The Black Youth Employment Crisis* (Richard B. Freeman and Harry Holzer, eds.); *Studies in State and Local Public Finance* (Harvey S.

Rosen, ed.); *The American Business Cycle: Continuity and Change* (Robert J. Gordon, ed.); *Economic Adjustment and Exchange Rates in Developing Countries* (Sebastian Edwards and Liaquat Ahamed, eds.); *Long-Term Factors in American Economic Growth* (Stanley L. Engerman and Robert E. Gallman, eds.). In addition, the *NBER Macroeconomics Annual*, edited by Stanley Fischer, began publication.

In addition to this new annual volume, the NBER plans to publish the following books in 1987: *Issues in Pension Economics* (Zvi Bodie, John B. Shoven, and David A. Wise, eds.); *Trade and Structural Change in Pacific Asia* (Colin I. Bradford and William H. Branson, eds.); *The Effects of Taxation on Capital Accumulation* (Martin Feldstein, ed.); *Public Sector Payrolls* (David A. Wise, ed.); *European–U.S. Trade Relations* (Robert E. Baldwin and André Sapir, eds.); *Money in Historical Perspective: Essays by Anna J. Schwartz* (Anna J. Schwartz); *International Aspects of Fiscal Policies* (Jacob A. Frenkel, ed.); *Taxes and Capital Formation* (Martin Feldstein, ed.); *The Economics of Tax Policy* (Lawrence H. Summers, ed.).

Periodicals. NBER publishes two periodicals, the *Digest* and the *Reporter*. The *Digest* provides summaries each month on recent NBER working papers of general interest. The quarterly *Reporter* contains longer summaries of recent program activity, reports of NBER conferences, reviews of recent work by NBER researchers, and abstracts of working papers issued during the previous quarter.

During 1986, Martin Feldstein continued as President of NBER and Geoffrey Carliner continued as Executive Director. Further information on NBER activities is available in the NBER *Reporter*, or from Geoffrey Carliner, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138.

DAVID KENDRICK, *Representative*

Report of the Committee on U.S.–China Exchanges in Economics

Exchanges in economic ideas and cooperation in economics education between the United States and the People's Republic of China entered a mature stage during 1986, as witnessed by the following events.

Responding to a request of Chinese Premier Zhao Zhiyang, I invited several AEA members including John Fei, Anthony Koo, and Lawrence Lau to work with leading officials of the State Commission on Restructuring the Economic System and the People's Bank on current issues of economic reform in China. We had a three-day meeting in Hong Kong in January 1986, and a five-day meeting in Peking in June 1986. Some of the issues discussed include price reform, reform of the administrative organization of state enterprises, reform of the banking system, macroeconomic control mechanisms, and problems of foreign trade and foreign investment. After the June meeting, I agreed with the Economic Restructuring Commission to organize two workshops a year to teach economics to national and provincial officials working on economic reform.

For the graduate economics program established at the People's University in Peking, I invited Leo Hurwicz and Elizabeth Li to teach in the spring semester, and Roger Gordon and Michelle White to teach in the fall semester of 1986. Some 48 students completed a one-year training program by July 1986.

A summer workshop on econometrics sponsored by the Chinese State Commission on Education took place from June 9 to July 19, 1986, serving the graduate students of the year-round program mentioned above and about 50 other graduate students and research economists from the People's Bank, the State Council's economic research institutes, the State Planning Commission, the State Statistics Bureau, and the Chinese Academy of Social Sciences. This was the third in a series of summer workshops to modernize economics education, following the micro- and macroeconomics workshops in 1984 and 1985, which I organized in cooperation with the Chinese Ministry of Education. Richard Quandt, Angus Deaton, Robert Engle, and I served as lecturers. This workshop and the graduate training program at the People's University received financial support from the Ford Foundation.

A group of eleven economists from the Research Institute on Economic Reform visited the United States. Eight of them joined a group of approximately twelve American and European economists in a conference on Chinese economic reform held in October in Harriman, New York, organized by Bruce Reynolds and sponsored by the *Journal of Comparative Economics* and Union College. Sixteen papers were presented and discussed.

In the fall of 1985, I helped place 63 students sponsored by the Chinese Ministry (now State Commission) of Education to some 50 American and Canadian Universities for graduate studies in economics. In the fall of 1986 about 50 more students were so placed. The students by and large have performed very well. Most of them are receiving financial support from the host universities, while a minority receive support from the Chinese government or the Ford Foundation.

Many economists and graduate students from China and the United States visited each other's countries in 1986. The study of modern economics is flourishing in China.

GREGORY C. CHOW, *Chair*

Report of the Committee on U.S.–Soviet Exchanges

The ninth U.S.–Soviet Economic Symposium was held at Tufts University, June 9–12, 1986, on the subject "Aspects of the Economics of Agriculture." The Soviet delegation consisted of nine economists, seven of whom presented papers. The American group consisted of six specialists on U.S. agriculture, each of whom presented a paper, and six experts on the Soviet economy, four of whom were also specialists on Soviet agriculture. Some of the topics covered were: urban-rural and farm-nonfarm ties and linkages; soil erosion; government intervention in and financing of agriculture; role of foreign trade in food and agricultural policies; and world food prospects to the year 2000. The American papers were excellent. The Soviet papers often contained interesting information, but were less analytical and suffered from hasty translation. (The U.S. delegation did not provide translation into Russian of their papers.) Most of the discussions were very interesting and quite frank, and it was with considerable difficulty that the time limits for each topic were observed.

The Soviet delegation, most of whom had never been in the United States, very much enjoyed their stay in the Boston area. One of the highlights was a visit to Wilson Farm, a large retail farm market in Lexington and by far the best in the area. The Russians could not get over the quality and variety of fruits and vegetables available. Many of them bought bananas. After leaving Boston, they spent some time in New York and Washington. In Washington, among other things, they visited a nearby farm and also spent an afternoon talking with U.S. specialists in the Soviet and East European Division of the U.S. Department of Agriculture.

The next symposium will be held in the USSR and is tentatively scheduled for the first week in September. Agricultural problems will again be the topic for discussion.

FRANKLYN D. HOLZMAN, *Chair*

Report of the Committee on the Status of Women in the Economics Profession

As in past years, this report contains both an assessment of trends in the status of women in the economics profession and a summary of the activities of the Committee.

The Changing Status of Women Economists. Last year I presented an analysis of trends in the status of women economists based on data from the Universal Academic Questionnaire. These data showed that an increasing proportion of women were obtaining B.A. and Ph.D. degrees in economics and were being hired at the Assistant Professor level. Progress into the higher ranks was less evident.

During the past year, the Committee followed a suggestion originally made by Alice Rivlin that we commission some more in-depth research on these trends. A research project has now been launched under the direction of Sue Berryman, a sociologist and Director of the new Center for Education and Work at Columbia University. The project is being funded by the Russell Sage Foundation and monitored by Alan Fechter, a member of the Committee. Alan has also helped us obtain access to data from the National Science Foundation's unpublished files and many of the preliminary tabulations provided here were graciously supplied by his staff.

The research project focuses on the career status of women with doctorates in economics relative to that of their male counterparts and on changes in that status over time. Specifically, it addresses five issues.

1) The mechanisms that underlie women's increasing shares of degrees in economics;

2) The nature and changes in the composition of the pool from which doctorates in economics ultimately emerge;

3) The nature and changes in the composition of successive cohorts of Ph.D. economists;

4) The dynamics that underlie new Ph.D. economists' entry into different employment sectors and activities; and

5) The extent to which and manner in which gender affects economists' careers within an employment sector and between sectors.

Although the analytical portions of the research are only now getting underway, and the findings will not be available until next year, we already have some interesting descriptive data on Ph.D. economists (see Tables 1 and 2).

Table 1 shows that over the past ten years, economists' salaries have risen substantially but have not kept pace with inflation, declining about 13 percent in real terms.[1] Men have done a little better than women, and women now earn about 86 percent as much as men.

Over this same period there has been a decline in the proportion of all economists employed in academia (and a corresponding rise in the proportion employed in business and industry) with the change for women being much sharper than the change for men.

Within the academic sector, a substantially higher proportion of faculty are tenured than was true ten years ago. Women have made gains relative to men, but are still considerably less likely to be tenured and much less likely to be full professors where there has been no progress in closing the gender gap.

Table 2 shows that women represent 15 percent of all new Ph.D.s awarded, up from 10 percent ten years ago.[2] The time it takes

[1] The data in Table 1 are from the *Survey of Doctorate Recipients*, a "rolling" longitudinal survey begun in 1973. It is a one-eighth sample of all U.S. citizens with doctorates in economics, all foreign-born economists with doctorates who are working in the United States, and all those with doctorates in areas other than economics who are working as economists.

[2] The data in Table 2 are from the *Doctorate Records File*, a cross-sectional survey administered annually since 1920 to the universe of new Ph.D.s from American universities.

TABLE 1—SELECTED STATISTICS ON PH.D. ECONOMISTS

	1975	1985	Percentage Change 1975–85[a]
Median Salary			
Male	$26,855	$46,740	74.0
Female	24,125	40,002	65.8
Total	26,721	46,300	73.3
Ratio (F/M)	0.90	0.86	−4.3
Proportion employed in Academia			
Male	0.71	0.67	−5.6
Female	0.73	0.58	−20.5
Total	0.71	0.66	−7.0
Ratio (F/M)	1.03	0.87	−15.8
Proportion of Academics Tenured			
Male	0.69	0.90	30.4
Female	0.47	0.78	66.0
Total	0.68	0.89	30.9
Ratio (F/M)	0.68	0.87	27.2
Proportion of Total Faculty who are Full Professors			
Male	0.47	0.50	6.4
Female	0.29	0.30	3.4
Total	0.46	0.47	2.2
Ratio (F/M)	0.62	0.60	−2.8

Source: Unpublished data from the National Science Foundation, *Survey of Doctorate Recipients*.

[a] Adjusting for inflation (using the CPI) presents a very different picture. The real median salary declined 12.9 percent for males, 17.0 percent for females, and 13.3 percent for all Ph.D. economists.

TABLE 2—SELECTED STATISTICS ON NEW PH.D. ECONOMISTS

	1975	1985	Percentage Change 1975–85
Ph.D. Degrees Produced			
Male	809	688	−15.0
Female	86	124	44.2
Total	895	812	−9.3
Percent Female (F/T)	9.6	15.2	58.3
Median Years from B.A. to Ph.D. for New Ph.D.s			
Male	7.6	8.9	17.1
Female	7.7	8.4	9.1
Total	7.6	8.8	15.8
Ratio (F/M)	1.01	0.94	−6.8
Percent of New Ph.D.s Planning to Enter Academia			
Male	64.6	63.0	−2.5
Female	78.9	49.0	−37.9
Total	65.9	60.8	−7.7
Ratio (F/M)	1.22	0.78	−36.3

Source: Unpublished data from the National Science Foundation, *Doctorate Records File*.

to move from B.A. to Ph.D. has crept up over this period and is a little higher for men than for women. This trend may be related to the declining availability of government financial support and the rising proportion of graduate training being financed out of a student's own or his/her family's resources. (The latter proportion increased from 25 percent in 1977 to 30 percent in 1985). Consistent with the data for all Ph.D. economists cited above, there has been a decline in the proportion of new Ph.D.s planning to enter academia, and most of this decline is concentrated among women. Whereas in 1975, almost four-fifths of women Ph.D.s planned an academic career, the proportion is now only half.

Figure 1 shows the age-earnings profile for men and women economists (using cross-sectional data for 1985). The biggest salary gaps occur in the prime years (ages 40–60). Further analysis should show the extent to which this is related to women's late entry into the profession, their scholarly productivity, discrimination, or other factors.

Committee Activities. The Committee has had an extremely busy year. In addition to holding three Committee meetings and sponsoring numerous events at regional and national meetings, we updated and published a new roster of women economists, mailed out three issues of our *Newsletter* (all on time!) and obtained funding for the new research described earlier. None of this would have been possible without the dedicated efforts of the members of our Committee. I particularly want to thank Nancy Gordon for taking on the chores of editing the newsletter and Joan Haworth for maintaining our mailing list and updating the roster.

A number of issues have occupied our attention this year including: (1) the focus and organization of CSWEP-sponsored sessions at the AEA meetings (we decided to experiment with a somewhat broader definition of gender-related issues); (2) the desirability of publishing a roster, now that an AEA directory comes out regularly (we decided to continue to maintain our own lists but to publish a roster less frequently and move toward more cooperative arrangements with the AEA in the future); (3) the need for

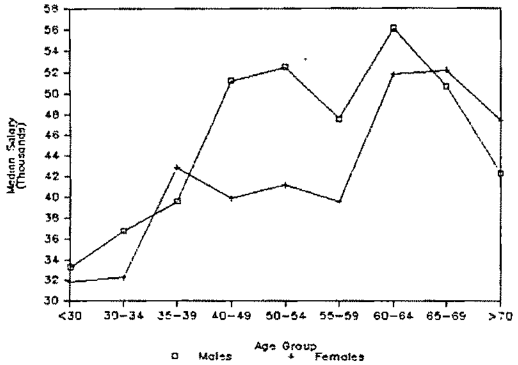


FIGURE 1. MEDIAN SALARIES OF PH.D. ECONOMISTS BY AGE AND GENDER, 1985

greater outreach, especially to students and younger members of the profession (a letter was sent out this year to our colleagues encouraging them to provide information about CSWEP to more junior faculty and students); and (4) the benefits and costs of blind-refereeing.

The last-mentioned issue surfaced very strongly at the open business meeting held in December 1985, and we published an article on the topic by Linda Edwards and Marianne Ferber in the fall issue of our *Newsletter*. As they note, research by psychologists has shown that people's assessment of the *same* manuscript is influenced by whether a male or female name appears on the cover, strongly suggesting that there is a bias against women authors. In addition, a study by Ferber and Michelle Teiman of 36 economics journals found that, when double-blind reviewing procedures were followed, articles authored by women were almost twice as likely to be accepted as articles authored by men, whereas when double-blind reviewing was not used, manuscripts submitted by men were somewhat more likely to be accepted. (As dramatic as these differences are, they were not statistically significant, primarily because of low sample sizes for women.)

Although some may not consider this evidence conclusive, the Committee believes that double-blind refereeing is strongly to be preferred as a matter of principle. It is likely to be perceived as fairer not only by women, but by less established members of the profession in general. As of 1986, of the 38 journals that provided information to Edwards and Ferber, only 14 practiced double-blind reviewing. We urge that all economics journals adopt this procedure and that the Executive Committee consider the appropriateness of current practices with respect to journals sponsored by the AEA.

Five members of the Committee completed their term this year: Lourdes Beneria, Bernadette Chachere, Mary Fish, Sharon Megdal, and Michelle White. I want to thank them for their service to the Committee and to make special note of the contributions of Fish, Megdal, and White who served as regional chairs and made major contributions to the Committee's work.

ISABEL V. SAWHILL, *Chair*

AMERICAN ECONOMIC ASSOCIATION

1987 ANNUAL MEMBERSHIP RATES

Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

● Regular members with annual incomes of $30,000 or less	$38.50
● Regular members with annual incomes above $30,000 but no more than $40,000	$46.20
● Regular members with annual incomes above $40,000	$53.90
● Junior members (available to registered students for three years only). Student status must be certified by your major professor or school registrar	$19.25
● In Countries other than the U.S.A., Add $12.00 to cover postage.	
● Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association)	$7.70

Please begin my issues with:

□ **March** □ **June** (Includes *Papers and Proceedings*) □ **September** □ **December**

First Name and Initial ____________ Last Name ____________ Suffix ____________

Address Line 1 ____________

Address Line 2 ____________

City ____________

State or Country ____________ Zip/Postal Code ____________

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

□ □

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) ____________

Below for Junior Members Only

I certify that the person named above is enrolled as a student at ____________

____

Authorized Signature ____________

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.

Computer Access to Articles in the JEL Subject Index

Online computer access to the *JEL* and *Index of Economic Articles* database of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 *(Economic Literature Index)* contains complete bibliographic citations to articles from the nearly 300 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts published in *JEL* since June 1984 are also available as part of the full bibliographic record. The *Economic Literature Index* also includes citations to articles in the 1979 and 1980 collective volumes (collected papers, proceedings, etc.) for the *Index* database; other years will be added as soon as completed. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers. (For a complete description of the *Economic Literature Index* with search examples and suggestions for searching techniques, see the article "Online Information Retrieval for Economists—The Economic Literature Index," in the December 1985 issue of the *Journal of Economic Literature.)*

Access Options:

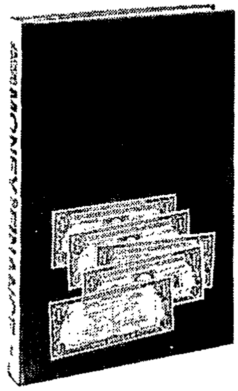
- **DIALOG** offers a variety of contract choices, including the option (for a low annual fee) to pay for only what you use. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call: DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800-3-DIALOG or 800-334-2564).

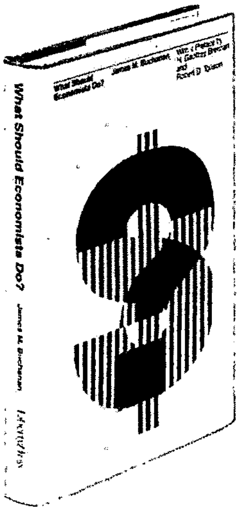
- **Knowledge Index,** a DIALOG service available after 6 p.m. and on weekends, may be accessed at the low rate of $24/hour, charged to a major credit card. A one time start-up fee of $35.00 buys 2 hours free time during the first month after log-on. Call 800-3-DIALOG for information.

- **EasyNet,** a gateway service, provides menus to guide the untrained user through database searches in DIALOG and other databases. For information, call 1-800-841-9553 or dial up **EasyNet** on your terminal (1-800-EASYNET) and pay for your search by credit card.

Classroom Instruction:

- DIALOG's Classroom Instruction Program, available at a special rate of $15/connect hour to academic institutions for supervised instruction, permits teachers to incorporate online bibliographic searching in their courses. For information, contact DIALOG or your librarian.





What Should Economists Do?
James M. Buchanan
LibertyPress

FELLOWSHIP PROGRAM FOR STUDIES RELATED TO THE KOREAN ECONOMY

The Korea Development Institute (KDI) offers fellowships to American Ph.D. candidates and faculty members in economics. Recipients will normally spend six months to one year in Seoul conducting research on subjects related to the Korean economy. It is our hope that high-caliber research on the processes and consequences of Korea's growth will greatly contribute to our understanding of the mechanisms of economic, political, and social change, thereby aiding future policy formation for Korea and for other developing countries.

Ph.D candidates will receive up to $15,000 per year for living expenses and tuition, plus one round-trip airfare to Korea. Faculty members will normally receive $2,000 a month, round-trip tickets for the fellow and spouse, and an apartment. Both Ph.D. and faculty fellows will be provided with office space and access to KDI's excellent library and computer facilities. Knowledge of the Korean language is not necessary as fellows will be provided with limited research assistance to deal with standard problems (e.g., translating table headers, but not long textual materials).

Applications from Ph.D candidates should include a resume, transcripts of undergraduate and graduate courses, a dissertation prospectus, two letters of recommendation (including one from the chairman of the dissertation committee), an annual budget estimate, and information on other scholarships which they hold. Applications will normally be reviewed twice each year, with deadlines at the end of February and August, and announcements of the awards made within two months.

Applications from faculty members should include a resume, a research proposal, and information on other funds available to them. They are accepted at any time.

All applications as well as requests for additional information, should be sent to Professor Leroy P. Jones, Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215. Applications will be screened by Professor Jones; Professor Lawrence Krause (U.C. San Diego Graduate School of International Relations and Pacific Studies); Professor Lawrence Lau (Stanford University); Professor Dwight H. Perkins (Harvard University); Professor Gustav Ranis (Yale University); and Dr. Park Yung Chul, President of KDI.

CRISIS ECONOMICS

ALLY VERSUS ALLY

America, Europe, and the Siberian Pipeline Crisis

Antony J. Blinken

With an alarming balance of trade in the US, and new leadership in the USSR interested in increasing trade with the West, issues that shook the Atlantic Alliance over the construction of the Siberian natural gas pipeline are being raised again. **Ally versus Ally** covers all aspects of the pipeline crisis, what happened and why, from both US and European perspectives. And, this lively and provocative book addresses the larger issue — how to develop a foreign policy that balances national security needs with economic ones.

A Praeger Publishers Title
paper: $12.95 (tent.) cloth: $29.95 (tent.) April 1987

THE STEEL CRISIS

The Economics and Politics of a Declining Industry

William Scheuerman

Why has the domestic steel industry collapsed? What are the political implications of a major industry in decline? And can the industry become viable again? **The Steel Crisis** responds to these, among other questions, and suggests that worker and community control of abandoned mills may be the answer.

A Praeger Publishers Title
cloth: $36.95 1986

Also of interest . . .

THE RECONSTRUCTION OF ECONOMICS

An Analysis of the Fundamentals of Institutional Economics

Allan G. Gruchy

In the only comprehensive analytical survey published to date on the institutionalists, Gruchy clarifies the increasing need for a "humanized" economics to meet the complex problems facing modern society.

A Greenwood Press Title
cloth: $32.95 (tent.) June 1987

DRAWING THE LINE ON NATURAL GAS REGULATION

The Harvard Study on the Future of Natural Gas

edited by **Joseph P. Kalt** and **Frank C. Schuller**

Prepared under the auspices of the Energy and Environmental Policy, Harvard University

A Quorum Books Title
cloth: $39.95 February 1987

WORLD HUNGER

A Neo-Malthusian Perspective

Mitchell Kellman

A Praeger Publishers Title
$38.95 February 1987

SUCCESSFUL CORPORATE TURNAROUNDS

Eugene F. Finkin

This book removes the veil of mystery surrounding successful turnaround management.

A Quorum Books Title
cloth: $35.00 May 1987

Praeger Publishers
Greenwood Press
Divisions of Greenwood Press, Inc.
88 Post Road West • P.O. Box 5007
Westport, CT 06881

MINIMAL RATIONALITY

Christopher Cherniak

Christopher Cherniak boldly challenges the myth of Man the Rational Animal and the central role of the "perfectly rational agent" in philosophy, psychology, economics, and cognitive science. He presents a more realistic theory based on the limits to rationality that can play a similar generative role in the human sciences, and seeks to determine the minimal rationality an actual agent must possess. $19.95

DOLLARS, DEBTS, AND DEFICITS

Rudiger Dornbusch

"Professor Dornbusch provides a clear, fascinating, timely and provocative analysis of the key policy issues faced by economists and policy makers concerned with the world economy."—Jacob A. Frenkel, David Rockefeller Professor of International Economics, University of Chicago $20.00

MIT Paperbacks

THE MANAGEMENT CHALLENGE

Japanese Views

edited by Lester C. Thurow

What works for the Japanese? As our nation's balance of trade with Japan dips eastward, American managers ask how Japan's economy has outperformed ours during the past 30 years. Some of the answers may be found in the contributions in this book which are unique in presenting Japanese management as the Japanese see it. They provide American managers with insights into practices that may spark new solutions to old economic problems. $7.95 paperback

"A number of game theory texts have appeared in recent years, but, measured by the comprehensive coverage of modern developments in game theory, none comes even close to Martin Shubik's two volumes."—Mordecai Kurz, *Journal of Political Economy*

A GAME-THEORETIC APPROACH TO POLITICAL ECONOMY

Martin Shubik

This companion volume to *Game Theory in the Social Sciences* sketches a unification of several branches of political economy on the basis of the theory of games. Its five parts cover Basics, Oligopoly, Cooperative Models of Closed Economic Systems, Strategic Models of Closed Economic Systems, and Externalities and Public Goods. $15.95 paperback

GAME THEORY IN THE SOCIAL SCIENCES

Concepts and Solutions

Martin Shubik $12.95 paperback

The MIT Press

55 Hayward Street, Cambridge MA 02142